

Kaufman - Trading Systems & Methods - Chap21

Chap21

VISIT...

LANZAROTE
Caliente.com

21 Testing

Testing a trading system was once a „cry simple process. Not so long ago it was difficult to get data, you needed to apply your rules manually to each price. you kept a handwritten record of your trades, and it took a long rime. You were careful not to begin unless you were reasonably sure that the method had a good chance of success; this was based on an understanding of the fundamentals or market experience.

During the past 20 years that process has changed, not always for the better. Those who have applied the same diligence and selectivity to the automation of systems are likely have gotten better results and more useful information than ever before. Others who have become careless because of the power and convenience of trading strategy developmerit software are not having much success. Bigger and faster computers do not mean better results.

Computers give back as much as you put in. The tools alone do not result in a successful trading program. The thought process and creativity, are needed to be competitive in today's market, although thoroughness and focus are qualities that can never be underestimated. Certainly, computers allow us to solve problems of magnitude and complexity that we could never have considered before, and do it in a wav that is easy to understand. We can manipulate incredible amounts of data and find out not only. whether our ideas are sound, but the risks associated with various events.

One of the most interesting techniques to develop from computerization is optimization. It is not the mathematical process of finding a local maximum in a field of continuous contours; instead, it is the iterative process of repeatedly testing data to find the single best moving average speed, pointandfigure box size. stoploss size. or other information. That is not to imply that it could not be more. Optimization is an irresistible method when a computer is available, and, in an indiscriminate way, it has replaced logical selection. it is frequently the means of creating strategies referred to as black box methods, and often results in fine tuning, which produces systems that invariably, generate high expectations but poor actual performance.

Computers present a serious, often futile, dilemma for the analy st. if a trendfollowing system is the objective and some form of smoothing technique is selected (a simple or stepweighted moving average, exponential smoothing). how do You select the right speed, Twenty years ago, the 5, 10, and 20day moving averages ,Nerc most popular because they represented a 1week, 2week, and 1month time interval. They. were also easy to calculate. In fact, the 10day moving average was the most common because no division was necessary to arrive at the average. Unfortunately, that simple approach no longer works.

Consider what is necessary to create a trading strategy. For a trend sy stem. once the trend period has been logically selected (e.g., 10day.) and the trading rules determined (such as a stoploss), you will want to know the performance of that choice over past markets. If it proves successful, you may trade using that method: however, if it does not nicer expectations

using the test data, what do you do? With a computer, the answer is easy: try another trend speed and see if the results are better. This follows a perfectly natural progression: after all, who would use a system today that has failed to be profitable in the past,

504

It is obvious that a trader with a talent for computers who has purchased a TradeStation, MetaStock, or another software package for testing, will take an organized approach to experimenting with a broad selection of trend speeds, box sizes, stoploss points, and trading rules to arrive at a system with an excellent trading profile. When no satisfactory combination can be found for a specific market, that market will not be traded. This method of optimization has become a dominant factor in the development of trading systems; it is entirely dependent on having a computer.

The following sections will discuss various aspects of testing and optimization. Performed properly, testing can teach a great deal; done incorrectly, it can be misleading and completely illogical. It is not always easy to know which case has just been satisfied.

EXPECTATIONS

Expectations are an underrated part of system development. It forces you to define your ideas in advance and put substance into those plans. You need to decide the rules, time period over which the system should work, the relative risk, the proportion of profitable trades, and other characteristics. The importance of this comes when you see actual test results.

Testing should be the process of validating your ideas. That first requires that you define those ideas in a clear plan and decide whether it should work for hourly, daily, or weekly intervals. If the test results confirm your ideas, then you can have confidence in the strategy. If they differ, you know that something is wrong, either with your plan or the way it was entered into the computer. Without expectations, there can be no validation or confirmation; therefore, if you indiscriminately test all indicators combined together, you have no idea whether you have found a good idea or simply overfit the data.

Setting Your Objective

There are a number of steps that must be carefully performed before testing can begin. These steps are more important than the actual testing and will decide what you are looking for and how you will use the results. Correctly defined, the final system will have realistic goals and predictive qualities. Incorrectly done, it will look successful but fail in actual trading.

First, define the test objective. What results should the test present so that you can determine its success? Is it the highest profits possible from the test strategy, the frequency of profitable trades, or the average profit to average loss ratio? Testing software gives you a choice, but doesn't allow a combination of these items; most likely, it will be a combination that you want. For example, if maximum profits are used as the performance criteria, which is most often the default case, the resulting system may have

one or two large profits (as Iraq's invasion of Kuwait in 1989 and the U.S. retaliation in January 1991) and an overwhelming number of small losses before and after it. The same performance would also show a very impressive average profit to average loss ratio. The highest frequency of profitable trades can also be inadequate if there are very small profits and large losses.

A popular way to view results is to look for a test profile with high profits and low risk. If we simply choose the largest equity drop as a measure of risk and net profits as a reward, a measurement function for best results could be

Test function = net profits x (100 / percentage equity drop)

which will reduce profits according to the size of the equity drop. The larger the losses, the smaller the test function value. A higher test function value should yield a smoother performance.

505

IDENTIFYING THE PARAMETERS

Once the test strategy is known, the parameters to be tested must be identified. A parameter is a value within the strategy that can be changed to vary the timing of the system. For example, parameters include:

1. Moving average speed (in hours, days, weeks)
2. Exponential smoothing constant (in percentage)
3. Band width around a moving average (number of standard deviations)
4. Stoploss values (in points or percentages)
5. Size of a point and figure box (in points)

The simplest optimization would be a test of the number of days in a moving average system, where all other values are fixed. The test would then simulate the results of the trading strategy by stepping through the possible range of moving average days: 1, 2, 3, and so forth, until a satisfactory range has been covered. This is called a 1-dimensional optimization.

Most systems have more than one important parameter. In the moving average system that is being used as an example, both the moving average period and the stoploss value are important. One test procedure selects the first moving average value, then tests all of the stoploss values; the second moving average value is then set and all of the stoploss values are retested. The process is repeated until all of the moving average values have been tested. This is a 2-dimensional optimization.

Distribution of Values to Be Tested

When more than two parameters are tested, the test time for the optimization may increase dramatically. For example, a test of 100 moving average values, 20 combinations of entry, exit bands, and 20 stoploss values gives $100 \times 20 \times 20 = 40,000$ tests. Even with today's faster computers, that's going to take more than 50 hours if each test takes 5 seconds. It may be necessary to select fewer values to test for each parameter and choose them more carefully. For a moving average trend, it is best to use all of the smaller values for the number of days and then fewer values as the numbers become larger. For example,

1, 2, 3, 4, 5, 6, 8, 10, 12, 15, 20, 25, 30, 40, 50

would be a better set of values than using all numbers between 1 and 50. If you observe that the difference between a 49 and 50day period is only 2%, while the difference between the 2 and 3period test is 50%, you realize that the set of tests at the two ends of the test series represents very different situations. If you expect to judge overall performance by the average of all tests, then using equal increments will weight this set of tests heavily toward the long end. By reducing the tests using percentage increments, you not only improve the test time but you improve the results. The moving average example shows that it is not possible to equalize the percentages due to the restriction of integer values at the low end, but by replacing this method with a smoothing constant used in an exponential smoothing, the problem can be fixed. The need to distribute results evenly over a set of tests will be important for comparing two systems and finding robust results. topics discussed later in this chapter.

Types of Test Variables

Test variables can be of three forms: continuous, discrete, or coded (alphabetic). To interpret and display test results properly, it is important to identify these forms in advance.

Continuous parameters refer to values, such as percentages, which can take on any rational number within a welldefined range. if a stoploss level is defined as a per

506

centage, it may be tested beginning with the fractional value .02% and increasing in steps of .005% until 2.0% is reached.

Discrete parameters are whole numbers, or integer values, such as the number of days in a moving average.

Coded parameters represent a category of operations, also called a regime. For example, when the parameter value is A, a single moving average is used; when the value is B, a double moving average system is tested; and when the value is C, a moving average and a breakout are used.

It is important to distinguish between the first two parameter types, continuous and discrete, and the last one. The analyst can expect some pattern in the results when testing parameters that take on progressively larger or smaller values. The coded parameter, however, usually causes rule changes. There is no reason why a change of rules, which causes the system to switch from one regime to another, would result in any performance pattern that makes sense across these regimes. The first rule may be profitable, the second losing, and the third profitable. The display of results, discussed later, is only valid for continuous and discrete parameters tested in an incrementally ascending or descending manner.

SELECTING THE TEST DATA

Testing a trading strategy on a computer is different from verifying a charting technique or taking prices from the Wall Street Journal to check results manually. If a computer is to be used for testing, there must be a complete database of price history and, if more ambitious recently, of economic statistics as well. This is readily available from numerous vendors and normally includes daily price data on U.S. and major European

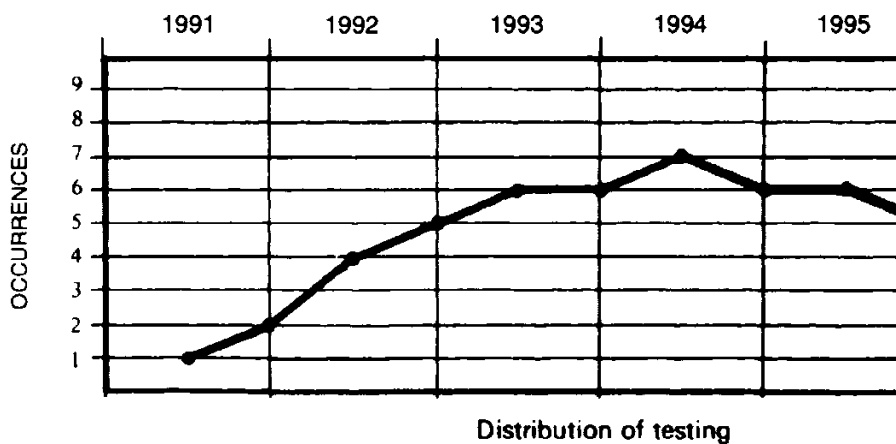
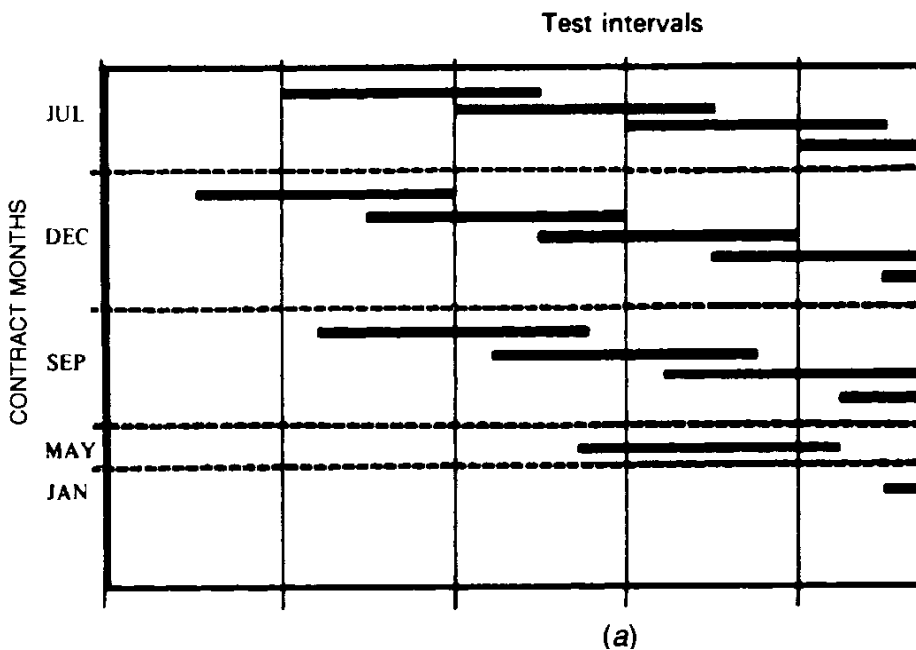
markets. When you purchase software to test your strategies, a database of daily prices is usually included. intraday (tick) data is limited to only a few vendors, the most wellknown being CQG and Tick Data, Inc. A complete database is an important asset. It should be kept current by automatic downloading at the end of each day.

Some computer systems, which are designed for trading strategies, have created special continuous test series specifically for optimization. While stock data or cash market prices are always continuous (except, of course, for stock splits), futures and options data present problems resulting from their limited contract life. To make this easier for testing, some vendors provide special features for combining shorter data periods into a single series. This gives the user the following choices:

- 1 . Individual full contracts. The entire futures contract, or more than one contract, are tested. Using full contracts is a tedious process and can result in overlapping test periods and duplications of results. Individual tables of results must be combined in a spreadsheet to be evaluated. Figure 2 11 gives an example of the test bias resulting from carelessly using all contracts within the range December 1992 through December 1996. Once the number of tests being performed at any one time is seen, it is easy to construct a distribution (Figure 2 1 1b) by counting the horizontal bars at specific points. This distribution will show the testing bias due to duplication. Because of the 18month life of the contracts, there was only 1 delivery month tested in the last half of 1991, compared with six contracts tested during the middle period, failing off to two contracts at the end of 1996.
2. Selected parts of contracts. Segments can be selected that include the last 30 to 90 days of a contract. By limiting a test to the period you are planning to trade, you have a good representation of expected performance, although there are a lot of little pieces to evaluate. When operating this way, you must remember to allow a windup period for your trend calculations prior to when you begin taking trading signals.

507

FIGURE211 Testing using full contracts from December 1992 through December 1996. (a) Test intervals. (b) Distribution of testing.



3. Backward (or forward) adjusted data. The data segments in item 2 are combined on a specific roll date by adjusting all the prices in that one contract by the difference between the closing prices of the new and old contract on the roll date. In a backward adjustment, all data prior to the adjustment date are changed by the differential on the roll date. This technique has become popular and works well for trend systems. The adjustments will cause large distortions in the oldest data, especially if the data go back more than 10 years. In the

extreme case, the adjusted price can even become negative. If the strategy uses a percentage value, this series will not give the right numbers; or, if you expect to compare various high and low prices over time, it is best to use another type of series.

4. Indexed data series. A continuous series is constructed by adding the price change each day to the prior index value. When there is a change of contract, the price change is that of the new contract compared with its prior price. The conversion into an indexed series is useful for comparing two markets and for con

508

structing portfolios, but has most of the same problems as its raw counterpart in item 3.

5. Constant forward series. In the same manner as the London metals contracts are quoted (ndays out), an artificial price is calculated by interpolating the prices of the delivery months on both sides of the date needed for this price series. This type of data series has also been called a perpetual contract and presents its own set of problems. Deferred contracts are not as volatile as the nearby contract, and formulated values create a series that is much smoother than the one you are likely to trade. This means less risk, probably less profits, and greater reliability in identifying trends. More important, you would be testing your ideas on a data series in which none of the prices actually occurred.

if all five methods have problems, which is best? If you use a trend calculation and do not rely on a percentage or compare the absolute levels of past prices, then you can use backward adjusted data (item 3). Although it requires more work, the safest choice is always to use the individual contracts, provided you take care not to duplicate performance over the same dates in different contracts unless you plan to trade that way.

In a seasonal or cyclic commodity, there is good reason for selecting more than one delivery month during each yearnew and old crops or production cycles can cause substantially unique price patterns during the same time period for different deferred deliveries. Even the proper overlapping of a traditionally nonseasonal product may pick up deviations due to anticipation or extreme shortterm demand or surplus. But the market shown in Figure 211a was not seasonal or cyclic, and the 18month contract duration caused the overlapping time intervals that simply distorted the test results.

SEARCHING FOR THE OPTIMAL RESULT

There are many complex techniques for optimizing; the two most basic will be discussed here. The first, and the only simple method, is to test all combinations of parameters. A single moving average system with a stoploss might have the following tests:

1. All moving average days from 1 to 50
2. Stoploss values from 0 to \$ 1,000 (or equivalent in points), in steps of \$ 10

The total number of tests would then be $50 \times 100 = 5,000$. When three parameters are tested, the number of tests increase rapidly. Taking

care to use the best parameter distribution, as discussed earlier, is a simple way to reduce the test time and improve the usefulness of the results. When there are too many tests, even with selected values, a wider spacing of values can be used to locate the general pattern of performance. Once the approximate range of parameter values is known, more detailed retesting can be done.

Calculate the number of tests you will run before beginning an optimization. Although the time may not be a problem, your test software may not be set up to hold the number of test results that you are creating. All computers have limits. If you generate as many as 1 ' 000 combinations, remember that you will need to evaluate them using a spreadsheet. Large numbers of rows and columns, combined with numerous markets, can create an ummanageably large spreadsheet.

Sequential Testing

If you have created a system on a grand scale and the number of tests and the number of parameters are too great, then mathematical optimization offers some solutions in the form of testing procedures. One caveat is that they only apply to continuous results, that is, performance that increases and decreases somewhat smoothly based on parameter val

509

ues. You cannot use this method to test coded variables that cause an abrupt shift in the methodology. In the simplest example, the procedure is:

1. Test one parameter, such as the moving average period.
2. Find the best result, and fix that value for the remaining tests.
3. Select the next parameter and test.
4. Go back to step 2.

This is called a sequential optimization. It has the advantage of reducing a 3parameter test case from $n_1 \times n_2 \times n_3$ total tests to $n_1 + n_2 + n_3$. It has the disadvantage, however, of not always finding the best combination of parameter values when the test is very large. Figure 212 shows a 2dimensional map of test results, a good way of visualizing the relationships between parameter sets. By testing all the combinations, the best choice was seen to be a 20day moving average with a 30point stoploss, netting 400. Using sequential testing, beginning with the moving average days, the best result is 15 days. Fixing that value and testing the stoploss values, we find that the best stoploss is at 30 points. While this result is good, it is not the peak result nor the next best 350. If the tests begin with the stoploss value, it ends at the same point; there, the best results cannot be found, whether the test begins with the rows or columns. Even though the area of greatest profitability is clear, the best choices in the first row and column do not peak at values consistent with the overall best result.

There are two ways that traditionally avoid this problem. One method is to list the parameters in order of importance, that is, those that would have the greatest impact on profitability, although you might choose this by its fundamental importance to the strategy. In this example, the number of days in the moving average is most likely to be the primary parameter. This doesn't work in Figure 212; however, mathematicians are prepared for this problem.

The final alternative is to choose a series of random starting values for each parameter and proceed to test a row or column beginning with those values. The object is to find the best results with as few of these tests as possible. Although each 3parameter test is only $n_1 + n_2 + n_3$, the total number of tests will accumulate steadily. When you have performed a number of searches beginning with a random seed, and found the same results,

FIGURE 212 Visualizing the results of a 2dimensional optimization.

		Moving average days				
		5	10	15	20	25
Stop-loss (points)	10	100	150	200	150	150
	20	200	150	250	300	300
	30	250	200	300	400	350
	40	200	150	200	300	300
	50	100	100	150	250	250

510

you can conclude with some certainty that you have located the peak. For example, if you begin searching in Figure 212 in four places, one row and one column inside the four corners (moving average of 10, stop of 20, 20 x 20, 10 x 40, and 20 x 40), you will always find the peak value. For a test of this size, if these were random seeds, three values that gave the same results would be enough. For larger tests, more seed values would be needed.

More on Test Continuity

Any system can be tested using a computer, but automatic selection of the best results only makes sense when they are continuous. A moving average system with stops is continuous in all directions. Adding a third dimension, an entry or exit delay (0, 1, or 2 days) will produce continuous results in the three directions. But not all tests satisfy this principle. It has already been briefly mentioned that rule changes are the most likely cause of discontinuity. Consider a test in which one variable is the speed of the trend and the other is a set of rules, L, S, and B, where L only takes long positions, S only takes short positions, and B takes both long and short positions. Although there may be interesting patterns and relationships,

averaging these results does not make sense.

If each parameter tested has values that result in a continuous profit/loss curve, there are mathematical techniques available to locate the point of maximum profit with the minimum steps. One method, optimization by steepest descent,' uses both the profit and the rate of change of profitability to determine the size of the steps taken during the optimization process. This method can be suboptimal, however, if there is more than one profit peak.

VISUALIZING AND INTERPRETING THE RESULTS

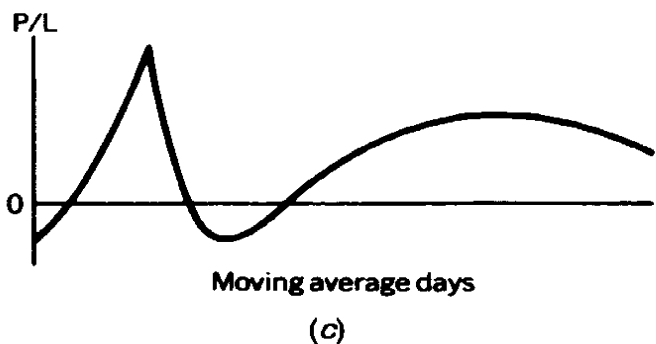
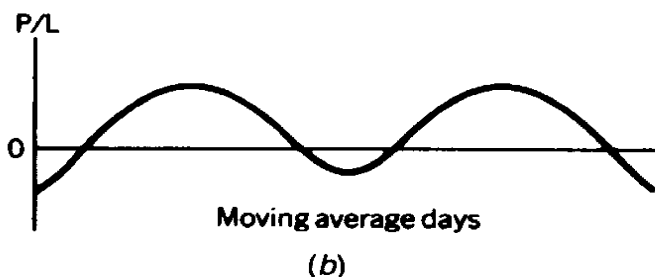
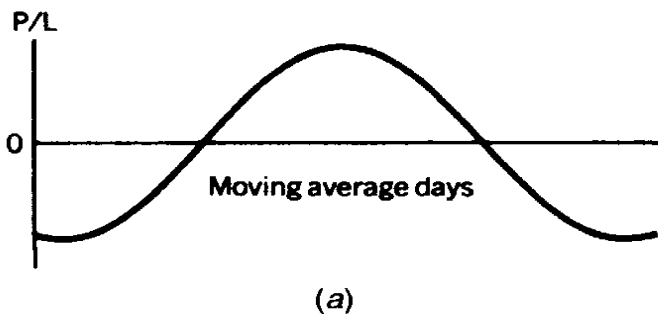
Visualizing the test results properly can make interpretation far easier. Figure 212 has already given a preview of how results might be presented in a way that shows continuity, rather than a simple list of one test result after the other. For example, if a 1parameter optimization is run in which all values are fixed except for the number of days in a moving average, then three possible results are shown in Figure 213.

In all three cases, the resulting profit or loss is shown on the left, and the number of moving average days is shown along the bottom. Figure 213a is a simple curve, indicating that both fast and slow parameters generate losses, although the midrange is profitable. it may be reasonable to select the number of moving average days that produced the highest profit due to the good continuity around that point.

Figure 213b shows two possible areas of success. Because both are about equal in size, there is no reason for choosing one over the other. One possibility is to trade one moving average from each peak. If you must choose only one, then it is most likely that the longerterm results are more stable, and considered a more conservative selection. A fast trading system can result in errors, poor execution prices, and large transaction costs. Unless these factors have been carefully incorporated into the trading strategy, a highly profitable simulation will result in large, real losses.

The preceding case is exaggerated in Figure 213c, where the fast system has much higher profits in a very narrow range, and the longterm trend has a wide range of success. The spike in the area of the 3day moving average is surrounded by losses, indicating that the high profits may have been caused by a shortlived price shock, or a pattern in the data that was perfect by chance.

FIGURE 213 Possible resulting patterns from a test that varies only the moving average days.



TwoParameter Tests

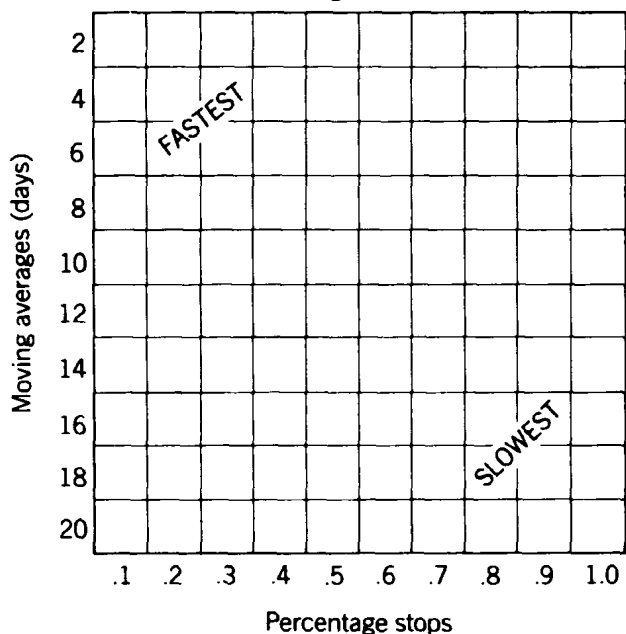
The most popular tests are those that have two parameters, either two trends of different periods or a trend and a stoploss value. Considering a moving average and stoploss, the best visualization is a grid (a 2dimensional table), with the moving average days along the left (rows) and the stoploss values at the bottom (columns). Figure 214 shows that the lowest number of days and the smallest stoploss value will give the profit/loss value in the top left corner box.

By presenting the results as shown in Figure 214, there is a continuity of performance in all directions. The upper left corner represents the fastest system—the one with the most trades; the bottom right corner shows the results of the slowest strategy. The patterns of the upper right may be similar to those of the lower left if the system displays some symmetry. A faster trend speed with a large stop (top right) and a slow trend with a small stop (bottom left) are likely to have the same number of trades and similar profitability.

Depending on the data used for testing as well as the trading rules, three patterns are most likely to appear in the results of a 2dimensional display. Figure 215a shows the simplest case of a single area of successful performance gradually tapering off. This is analogous to the singleparameter test shown in Figure 213a. Selecting the parameters that

512

FIGURE 214 Standard configuration for a 2dimensional optimization.



gave the center of the best performance area is the most reasonable, because moderate shifts in price patterns are still likely to be profitable.

The appearance of two areas of profitability are usually the results of a highly volatile market (Figure 215b). Prices that are moving quickly and sustain a major trend can be traded using a fast or slow model. The fast model is often more profitable because it captures more of the rising trend; it reacts faster and also profits from the decline. The slower trendfollowing approach captures less of the price move but keeps clear of the sharp reactions that stop out the middlespeed trends. The selection of parameters to use in real trading follows the same reasoning applied to Figures 213b and c. The third case, shown in Figure 215c, is one of erratic profits and

losses. The absence of a consistent pattern in the performance indicates that the trading strategy does not apply to the data.

Alternate Ways of Visualizing Results

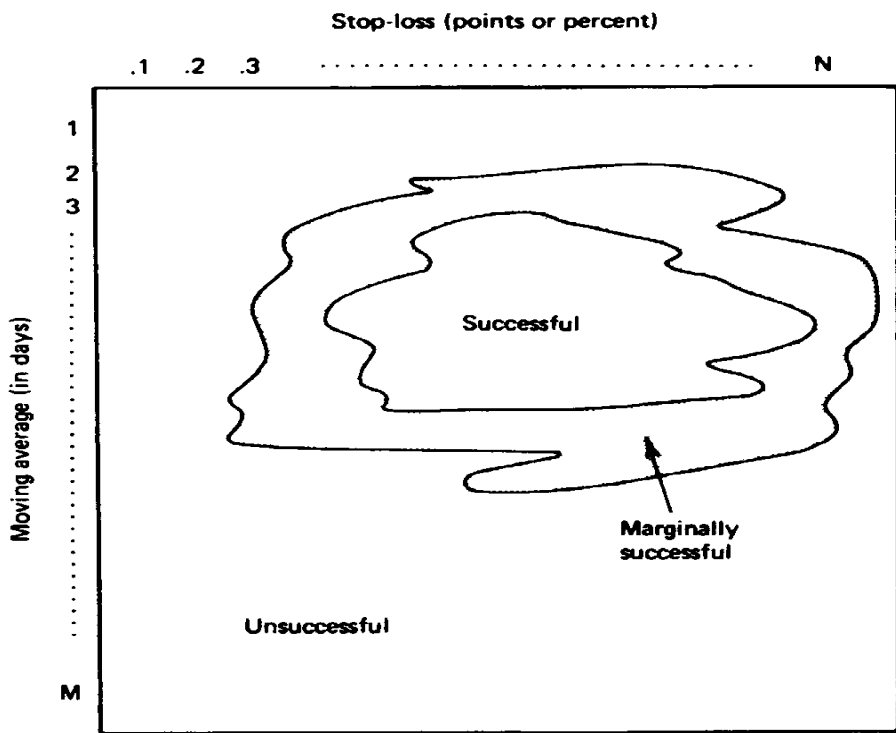
With software such as MathCad, you can see the 2parameter test as a contour or topological relief map. For some users this form can make it much easier to see the peaks and valleys of performance and is highly recommended. Any spreadsheet program will allow a surface plot, which will give you a rough look at the continuity of a 2parameter test, as shown in Figure 216. This 3dimensional graph gives the net profits of a moving average crossover system applied to the Swiss franc. The slower trend is shown along the front scale and the faster one along the right. There is a ridge running from the front to the back, indicating that a slow trend of about 50 to 60 days dominates the results. Other than the fastest trends of 3 days, it is easy to see that there are many combinations of fast and slow moving averages that are historically profitable.

Visualizing with Scatter Diagrams

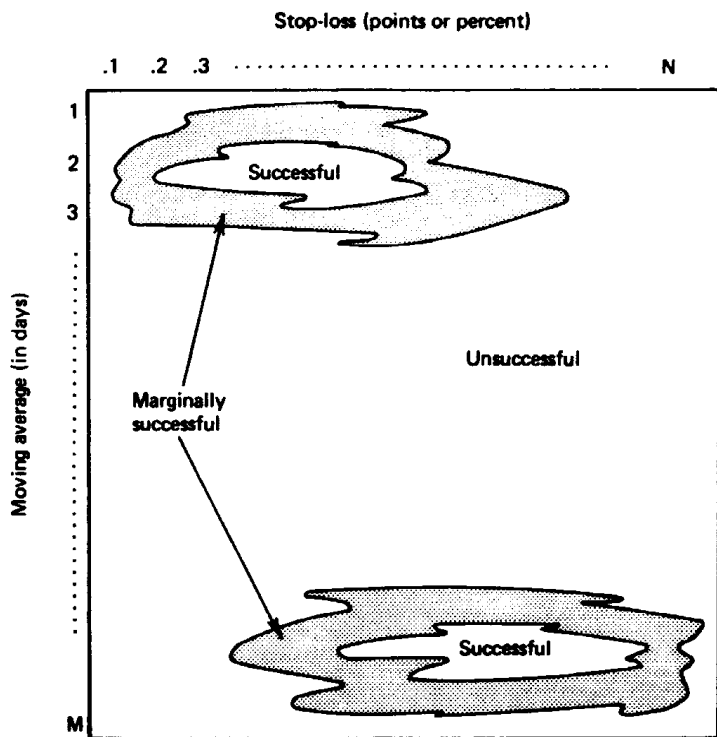
To see the performance of more than two variables, you can plot the results of all tests against each parameter (see Figure 217) using a scatter diagram (also called an XYPlot).

513

FIGURE 215 Patterns resulting from a 2dimensional optimization. (a) Single profitable area. (b) Two distinct profitable areas. (c) No obvious pattern.



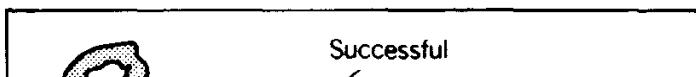
(a)

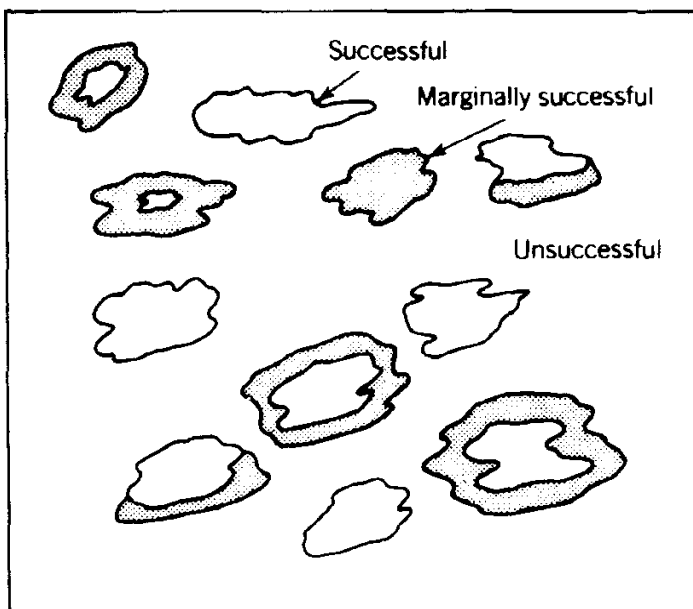


(b)

514

FIGURE215 (Continued)





(c)

This will give you a better understanding of how each value affects the total picture and allow you a way to visualize the robustness of each variable Using a Invariable example. a range of fast and slow moving averages were tested for the Swiss franc based on a crossover strategy. The results of the slow moving averages are plotted for all combinations of the fast moving average in Figure 217; the fast average is shown in Figure 21,7b.

The individual test results are shown as circles above the slow or fast moving average period in Figure 217. For the slow moving average, it is clear that profits are greatest when 50 and 60 days are used for the calculation period; however, it is not as clear which of those two values is the best choice. The 60day moving average posted profits higher than the 50day average, but also showed profits below the lowest 50day returns. A closer look shows that the average returns of the moving averages over those periods is about the same, but the 60day moving average has a higher variance. Normally, the best choice is the result that is most stable (the 50day calculation). although in this example we must be cautious that the small number of tests do not make the 50day, results look stable just by chance.

The results of the faster moving average, seen in Figure 2 1 b, show much greater variance. The 7period results have a very wide range including the highest returns however. the averages of the 7th through 15th periods are very similar. With variance in mind. the 13day calculation seems to be the best candidate. We can be confident that this method is robust, because nearly all combinations of fast and slow moving averages

gave profitable returns, net of a \$50 commission.

Standardizing Test Results

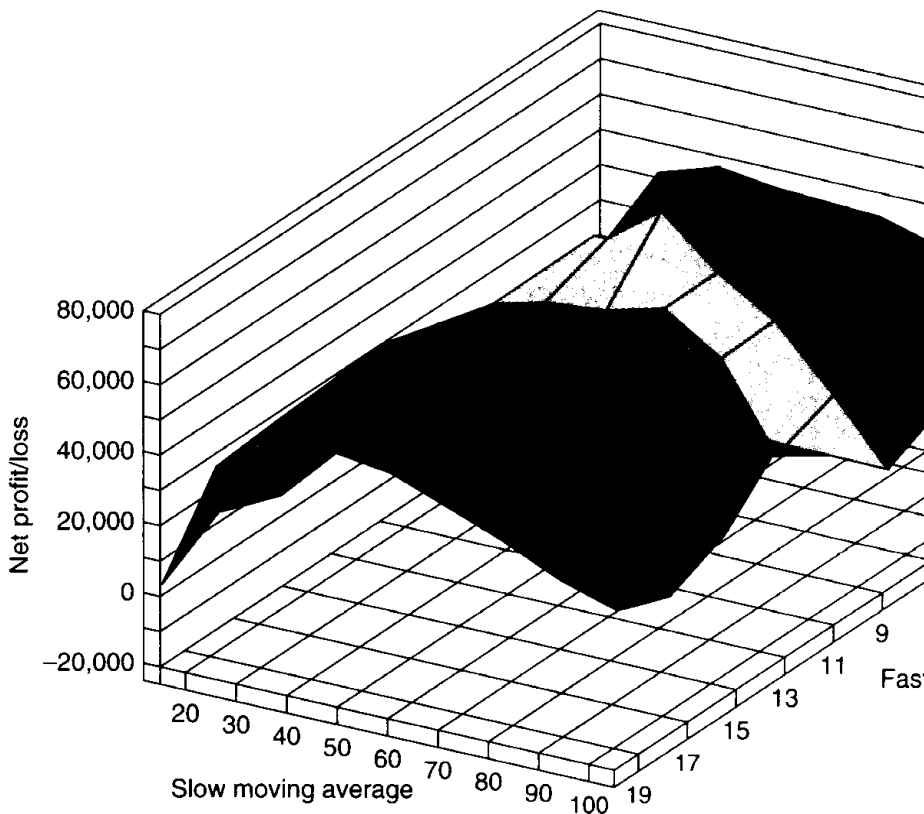
When testing a single system, the results of one set of parameters can easily be compared with other sets because the test period, commissions, and other basic values are all the same. However, over time you will test many different systems and variations of those systems, and you will want to compare them to decide which strategy is best. This requires some advanced planning.

Standardizing test results is the best way to increase the usefulness of extensive testing. This should include:

515

FIGURE 216 Threedimensional map of crossover model.

IMM Swiss franc, 1986–1996



1. Annualizing all values. Tests that you perform 6 months or 1 year apart are likely to use different time periods. Annualizing the results will make comparisons easier.
2. Risk adjusting. A profit of \$10,000 with a drawdown of \$5,000 is effectively the same as a profit of \$20,000 and a drawdown of \$10,000, given adequate reserves. Expressing the results as profits

divided by risk can be useful.

3. Adjusting for standard error The important difference between two tests that yield the same results, one with 2 trades and one with 50 trades, is that the one with only 2 trades is a less reliable result. If both posted 50% profitable trades, another losing trade would make the reliability of the second test drop to 49%, while the first case would drop to 33%. Instead, results should be presented at the same confidence level, which means subtracting the standard error from the current result. Analysts developing relatively slow trading systems will find that this procedure has a startling effect on expectations.

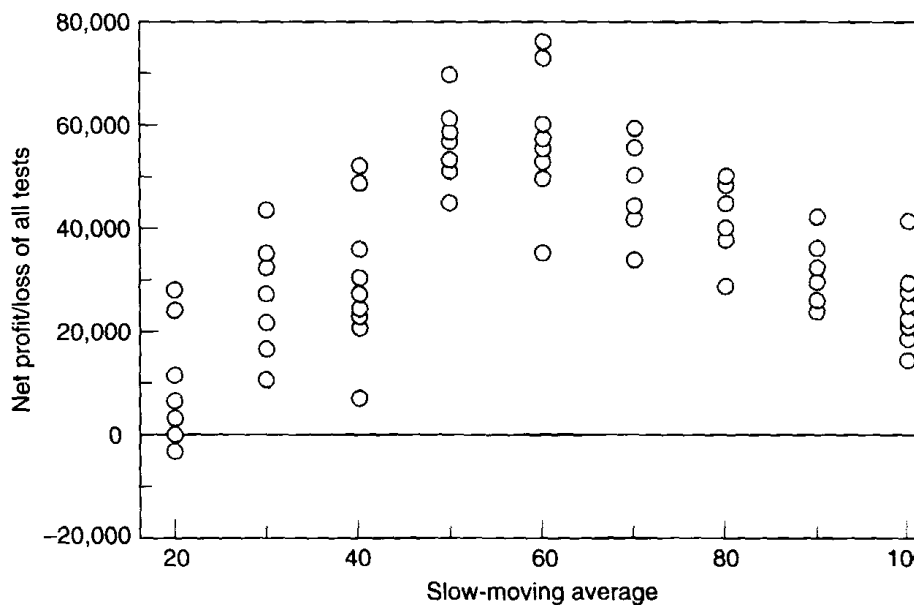
Averaging the Results

Because the testing process is computer dependent, it is desirable to make the selection of the best parameters automatic. This can usually be done effectively by averaging either the profits or other results displayed in the test map. In the case of a moving average system, first determine how broad the area of success must be so that it is not a spike, an anomaly in the results. For example, the results of a 3, 4, 5, and 6day test show profits of 1,000, 8,000, 3,000, and 4,000, respectively. The 4day case is clearly a profit spike and could be minimized by replacing its value with the average of the three other values or by the interpolated value of the two adjacent points (Figure 218). In general, the substitution process in which PLi replaces PLi,

516

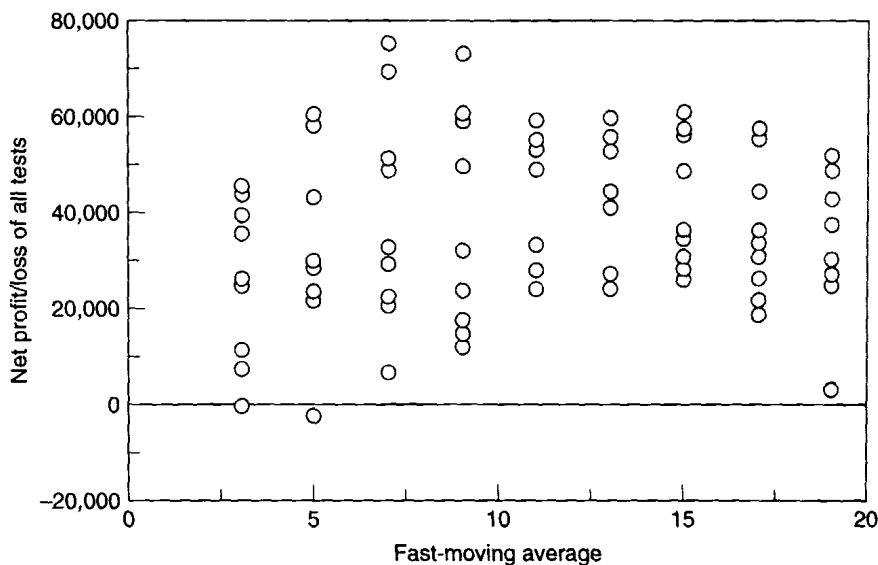
FIGURE217 Visualizing the performance of multiple parameters. (a) The slow moving aver. age peaks at 5060 days. (b) The fast average may be best at any value from 7 through 95.

Alternate display of 2-parameter test
 Swiss franc 1986–1996



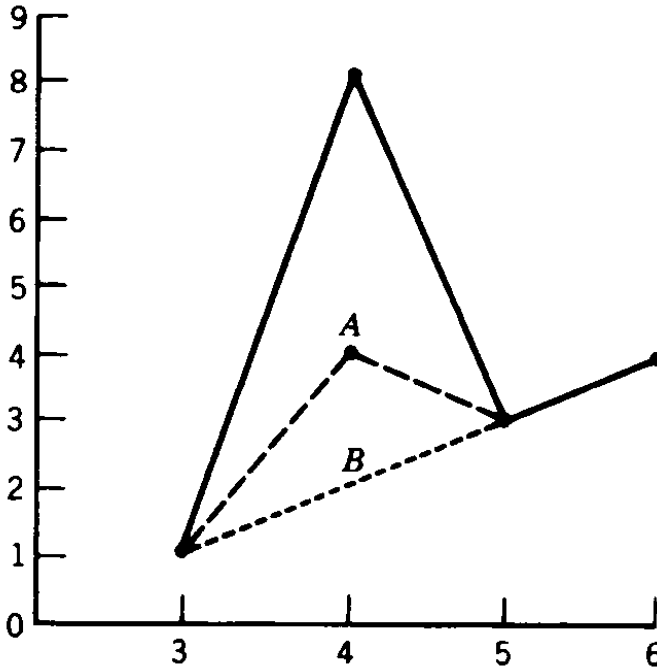
(a)

Alternate display of 2-parameter test
 IMM Swiss franc 1986–1996



(b)

FIGURE 218 Replacing a spike with the average A or an interpolation 8.



$$\overline{PL}_i = \frac{PL_{i-1} + PL_i + PL_{i+1}}{3}$$

will smooth the performance. The best parameter set is the highest value of PL, once all averages have been substituted for the raw test results.

it is more likely that a spike or erratic results will appear in areas that test shorter periods than in those that test longer trends. The difference between the results of a 2 and 3day moving average is greater than that of a 30 or 31day average or even a 30 or 40day test. If the optimization was scaled properly, leaving more space between tests of higher value in the manner of a percent difference, the simple averaging technique will be a good substitute for testing every value.

TwoParameter Averaging.

The results of a 2parameter test may also be averaged. Using the form discussed earlier (Figure 219), the results of each test will be denoted as PL_{ij} , the profit/loss associated with row i and column j of a 2dimensional display. The object is to replace each PL_{ij} with $\overline{PL}_{ij}$ the average value of its surroundings. This is done, as seen in Figure 219a, by taking an average of the eight test results adjacent to the ij th value as well as the center value. There are special cases when the ij th result is not fully surrounded but on the perimeter of the test map. Figure 219bd shows the

averaging technique used when the ij th box is a top, side, or corner value. The 9box average shown here is comparable to the 3point average in the 1parameter test. If the test has relatively small increments between calculation values and more test cases, an average of a larger area would be appropriate.

STEPFORWARD TESTING AND OUTOFSAMPLE DATA

The correct procedure for testing always includes reserving some data to be used afterward to validate the results. The data used during testing is called insample data, and the reserved portion for validation is called outofsample data. This concept should be familiar to everyone involved in the design or development of a system.

Traditionally, analysts use as much data as possible to determine whether their proposed strategy is sound. The use of long test periods assures a good sample of price patterns, including long periods of sideways movement, bull and bear markets, and a good

518

FIGURE219 Averaging of map output results. (a) Center average (9box). (b)Topedge average (6box). (c) Leftedge average (6box). (d) Upper left corner average (4box).

$PL_{i-1,j-1}$	$PL_{i-1,j}$	$PL_{i-1,j+1}$
$PL_{i,j-1}$	$PL_{i,j}$	$PL_{i,j+1}$
$PL_{i+1,j-1}$	$PL_{i+1,j}$	$PL_{i+1,j+1}$



(a)

$PL_{1,j-1}$	$PL_{1,j}$	$PL_{1,j+1}$
$PL_{2,j-1}$	$PL_{2,j}$	$PL_{2,j+1}$



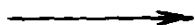
(b)

$PL_{i-1,1}$	$PL_{i-1,2}$
$PL_{i,1}$	$PL_{i,2}$
$PL_{i+1,1}$	$PL_{i+1,2}$



(c)

$PL_{1,1}$	$PL_{1,2}$
$PL_{2,1}$	$PL_{2,2}$



(d)

number of price shocks of various sizes. When more data are used, it is most likely that the results will show larger profits and larger losses.

When there is, for example, 10 years of data available for testing, a standard approach would be to test the oldest 9 years, find the best set of parameters, then test the system on the most recent year using the selected values. If the results of the outofsample period are similar to the insample performance, the system is considered validated.

One serious problem occurs when the results of outofsample testing fail to perform as expected. The results are inspected closely to find how this period differed from the past, and if successful, a rule change is incorporated into the program and the insample data retested. This repeated use of outofsample data is called a feedback process. Unfortunately, there is no longer any outofsample data to be used for validation. The improved results are the product of fitting the data and you are left uncertain about the ability to produce profits in the future. This is a problem that can only be resolved by a broad view of robust testing procedures discussed throughout this chapter.

519

step forward testing procedure

The concept of choosing a parameter set from a test of insample data, then applying the results to outofsample data, has been developed into a test process called 'stepforward testing' (also walkforward testing). This process goes as follows:

1. Select the total test period, for example 10 years of daily data from 1988 through 1997.
2. Select the size of the individual test intervals, for example 2 years.
3. Begin testing with 1988 and 1989 data.
4. Select the best parameters from those results.
5. Find the performance of the next 6 months of data (the first half of 1990) using the parameters selected in step 4. Accumulate this outofsample performance throughout the test process.
6. If there is more data to test, move the test period forward by 6 months (the second test will be from the second half of 1988 through the first half of 1990) and test the data; otherwise, go to step 8.
7. Go to step 4.
8. The results of the stepforward test are the accumulated results of the individual outofsample tests in step 5.

This process clearly simulates the traditional approach to test design, using insample and outofsample data in its proper order. Unfortunately, it does not correct the problems of reusing the outofsample data, which transforms that step to another piece of insample data.

Stepforward testing can introduce a bias that favors faster trading models. Because the insample test has only 2 years of data in our example, a longterm trading model may only post a few trades during this period, and one of the longer trades may be interrupted by the end of the data window and discarded. This makes the results of the longerterm models unreliable. In addition, these slower trading approaches will often start a trade at the beginning of the test data window at a point that is abnormal, that is, one that would not occur if the data were continuous.

The problem of shortterm bias can be identified by the optimum

parameters switching from shortterm to longterm in successive test periods. For example, if you test moving averages from 5 to 50 days, you will find one period shows 10 days as the best followed by a period in which 50 days is best, followed again by a period in which 15 days is best. This erratic behavior is a sign that the testing specifications are too limited. It can be corrected by extending the test intervals from 2 years to 5 years and making the outofsample period 1 year. None of these solutions, however, correct for the use of outofsample data more than once. More on this problem can be found in the following section "Retesting Procedure."

CHANGING RULES

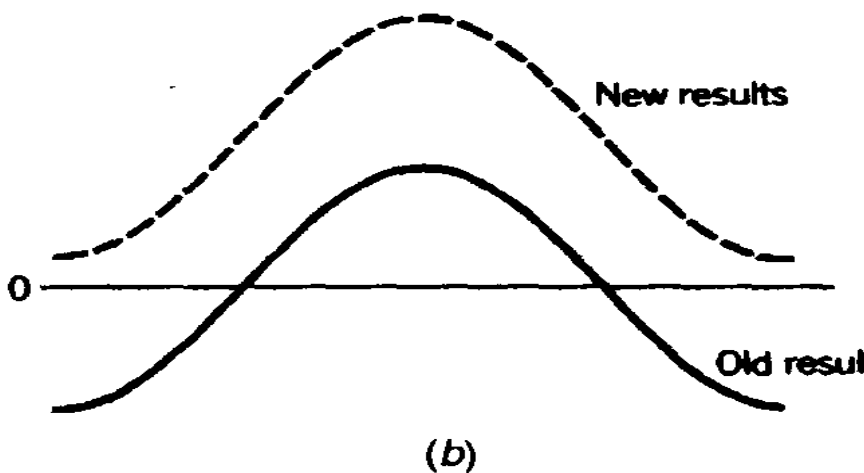
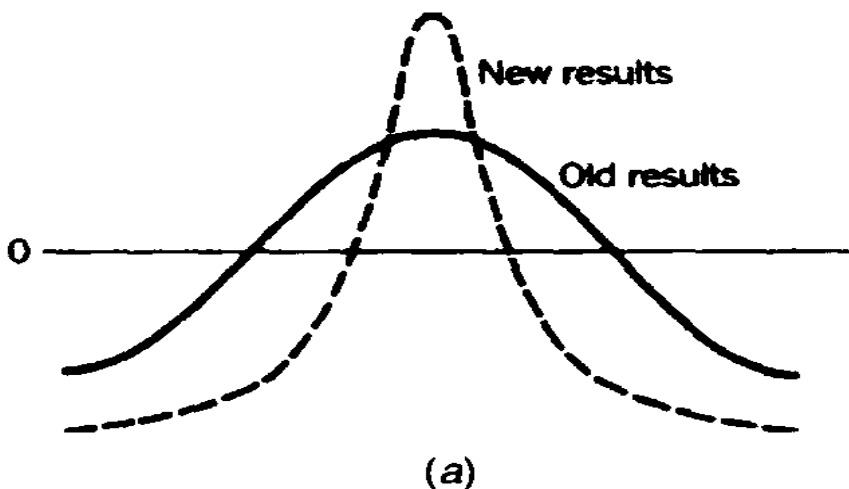
As testing progresses, it is inevitable that the analyst will want to modify a rule in the trading strategy. This is usually the result of inspecting the results of tests and noting that a specific pattern was not treated properly (or profitably). After some work the analyst introduces a special rule, which turns a previously losing situation into a profit. Figure 2110 shows two possible patterns in the test performance pattern based on this change.

if the rule change improves one price pattern at the cost of others, the complete test results will appear as shown in Figure 2110a. Higher profits at the previous peak and greater losses at the ends result in the same cumulative test profitability. Because the new

'A thorough presentation of both stepforward testing and optimum search techniques can be found in Robert Pardo, *Design, Testing and Optimization of Trading Systems* (John Wiley & Sons, 1992).

520

FIGURE 2110 Patterns resulting from changing rules. (a) A rule change that improves one situation at the cost of others. (b) A rule change resulting in general improvement.



rule caused the fitting of a specific pattern at the cost of added losses in other patterns, it should not be considered an improvement.

In Figure 2110b, the new rule improves performance in all cases. This type of pattern is desirable in optimizations. It is possible, of course, that the improvement was caused by a rule that corrects one case only in a way that is so specific that it does not affect any other trade. That type of rule fitting, on a sample of one pattern, is likely to harm results in the long run.

ARRIVING AT VALID TEST RESULTS

It is not unusual for the results of an optimization, especially one with many parameters, to appear perfect. All the trades could be profitable,

the equity drops might be small, and the strategy might perform in changing markets, and yet, the final system might have no real expectation of profit. To create a system with predictive qualities, rather than one that is historically fitted, requires preparation in advance of testing.

You Can't Prove a System (or Indicator) Works by One Case

You can always select a moving average or indicator gave a sell signal in the S&P just before the 1987 crash. That's not difficult when there are thousands (or millions) of combinations of trend speeds, filters, momentum indicators, divergence rules, and so on. However, what works in a specific case, and often an unusual one, is not likely to work in general. Manipulating the rules is only a waste of time. Showing that a combination of technical indicators would have profited from the last market crash has no bearing on the next drop.

The adage that "there are lies, half lies, and statistics" is too true. Statistics can prove a point or prove just the opposite, depending on how the numbers are presented. A mar

521

ket can be in an uptrend and a downtrend at the same time, based on the time interval over which you view it. You must be constantly aware of those problems that you cannot see the omissions.

It will be necessary for you to draw your own picture of system performance. "at is the chance of a big profit this year or this week? What is the worstcase scenario? Will the method survive a sharp change in market volatility? Can I choose a longterm or shortterm horizon with more confidence? Is an 8day moving average really better than a 10day breakout? The following guidelines will get you started by giving you one way to look at test results. From this you will be able to develop, over time, your own method of evaluation. As Jack Schwager has stated, "The mistake is extrapolating probable future performance on the basis of an isolated and wellchosen example from the past."

Searching for Robustness

A robust system is one that performs consistently in a wide variety of situations, including patterns that are still unforeseen. From a practical view, this translates into a method that has the fewest parameters and is tested over the most data. In a test by Futures Truth, the best performing systems commonly have four or less optimized variables.' The characteristics of a system with forecasting ability are:

1. It must be logical. Each rule and formula must be designed to capitalize on a real fundamental or price phenomenon. Discovering a price pattern or cycle through optimization may seem to be a revelation, but it is more likely to be an illusion. By testing enough patterns, it is statistically probable that one of them will seem to fit. Without a fundamental reason for the existence of that pattern, it is not safe to use it.
2. It must adjust to changing market conditions. A system that assumes that markets don't change, or that everything has been seen in past data, will suffer large equity swings. Selfadjusting features might include an inflator or deflator, stoploss values that change with volatility, and

rule shifts based on seasonal or nonseasonal years.

3. It must be tested properly. The principles of statistics state that the best tests use the most data. More data includes bull, bear, and sideways markets; large and small price shocks; and periods of instability and doldrums. There is no substitute for more data. Proper test procedures also suggest that you reserve some unseen data until the very end, so that you can validate your work on an outofsample period. If that is not available, you can always paper trade until you can compare the accumulated trading profile with the expectations defined by your tests.

Some analysts prefer a technique called blind simulation, commonly known as stepforward testing. In this procedure, discussed in a previous section, a series of fixedlength tests are defined and the parameters found to be optimal are used to trade using data in the next period forward. The process then moves forward by dropping off old data and adding an equal amount of new data. The results of this second test are again traded on forward data. The method is continued until the outofsample results include all of the available test data. This is an important concept and is also discussed in the section "Reoptimization" later in this chapter.

From time to time each of the three characteristics above will seem to be the most important. In the long run they must all be satisfied to have a robust program.

³Jack D. Schwager, *Schwager on Futures.. Technical Analysis* (John Wiley, & Sons, 1996, p. 673).

⁴John Hill, "Simple vs. complex," *Futures* (March 1996, p. 57).

522

Performance Criteria

Assuming that the proper principles have been followed, are the test results good enough? Although the strategy shows steady profits with a low equity drawdown, could a naive system have done better? Did the famous skeptic on television throw dam at buy and sell signals on a board and show a 75% return while your system only netted 25%

Benchmarks

It is necessary to have a benchmark that provides a way of measuring success. it is best if this is a welldocumented indicator, such as the S&P Index, the Lehman Brothers Treasury Index, a class of Fund Managers, or the Commodity Trading Advisor Index. Newspaper articles that highlight the spectacular profits or losses of one manager is not a good benchmark, but simply ex post selection (hindsight).

it may seem desirable to have achieved the recognition for having the largest gain during a single month, but you should focus on having very good gain over each year. The largest gain can only come with high risk. You should not be surprised if those investment advisor~s posting remarkably high returns in one month have had very erratic performance overall.

Measuring Test Results

A number of performance criteria are needed to evaluate any trading strategy during the test phase and to compare results under actual market conditions.

- 1 . Net profits or losses. Although not the most interesting statistic, the net

profit is the motivation for trading. While you would not select a system that produces a net loss, it is the other statistics that will tell which of the better results, if any, are realistic.

2. Number of trades. This simple value indicates whether your test was long enough to depend on the results. A few trades can appear very profitable, but the trading profile is not yet clear.
3. Percentage of profitable trades. Also called reliability, a high value of 60% tells you that the method captures profits regularly. A trend system will be working correctly if its reliability is near 40%.
4. Average net return per trade. With or without commissions and slippage costs, the average return per trade gives you an indication of how difficult it will be to realize the system returns. A theoretical average of \$50 per trade for a currency is likely to net less than one-half after slippage in normal markets, and when a U.S. economic report is released, slippage alone could be \$500 on a single trade.
5. Maximum drawdown. The largest equity swing from peak to valley, this measurement can be very erratic and is not likely to be the largest drawdown seen in the future; however, it gives you some idea of the minimum capital needed to trade this market. If the value is very small, it is likely to be based on a small amount of data, too many specific rules, or a narrow range of test parameters.
6. Annualized rate of return. The rate of return is the profit for a predetermined investment. The investment can be calculated in reverse by knowing the funds-at-risk to equity ratio, the maximum drawdown, or standard deviation of expected returns. These returns should be annualized to compare one test, market, or benchmark with another.
7. Total profits to total losses. This simple ratio gives a reasonable measure of the smoothness of the equity curve. As the ratio increases, the proportion of losses declines and the equity curve becomes smoother. Any value over 2.0 is very good.

Time to recovery. A large drawdown may be inevitable in a realistic system, but a shorter time to recovery is most desirable. A larger drawdown with a much faster recovery seems to be a better tradeoff for most investors.

523

9. Time in the market. All else being equal, a trading system that is in the market less than another system is preferable. If two systems have approximately the same returns and risk, then the one that has more time out of the market is actually exposed to less risk.
10. Smoothness of returns. In addition to the individual measurements described above, some form of traditional equity smoothness is desirable. This can be the standard deviation of the residuals when the linear regression of the returns is calculated, or it can be some weighted combination of the previous measurement. This measurement should reflect the type of equity profile you seek.

Having decided the measurements needed to evaluate the results of the testing, some other procedural issues must be resolved before testing

begins. This will help to avoid bending the rules to fit the results.

Defining the test range. Each optimization run should represent a broad sampling of tests over a wide range of parameter values. This range should be established in advance based on expectations of what is reasonable. If the returns in this range are not profitable, you must first understand why your ideas have not worked before looking for other periods in which the system might show profits. A thorough review of your trading rules may show that you have defined them wrong, not limited them to the best situations, or simply programmed them incorrectly.

Use the average of all test results. Most tests will show both profits and losses, with some areas of very attractive profits. If you tested 100 cases and 30 showed profits of about 30% per year, 30 showed breakeven results, and the last 40 showed various losses, you might say that the 30 profits represented a successful system. That assumes that the market will continue to perform in a way that allows those 30 parameter combinations to generate profits during the next year. It is better to assume that market changes will make it difficult to tell which combination of parameters will be the best in the future. Then you want a system that can generate profits with any parameter set, not just 30 out of 100. Perhaps that's too ambitious; nevertheless, one system is better than the other if the average of all tests is higher than the average of all tests in the other system. To be more precise, we could say that the best system is the one in which:

Adjusted Returns = Average of all tests - 1 standard deviation of all test
returns

is greatest. From the discussion of standard deviations in Chapter 2, we know that a return greater than the adjusted returns will occur 68% of the time.

Set aside data for final validation. To give yourself a way of knowing whether you have overfitted the data, do not use all the data in your tests. This is needed for outofsample verification at the end. If the results of using this data are not good, you will need to reassess the performance. If it is good, you have some assurance that your testing was done correctly, but there are no guarantees other than time.

Stability over time. Be sure that the test data was long enough to have included a substantial change in price, a major bull and bear market, and an extended sideways period at low volatility. By viewing performance over a number of smaller time intervals, instead of only one average or total figure, it is possible to assess the stability of the strategy.

Elimination of outlier periods. Realistically, a system should not base its profitability on a single major market move over the test interval. In particular, if the move that generated the profit was actually a price shock, the system had only a 50% chance of posting that profit, and has an equal chance of a loss in the future. It may be practical to remove the profits generated on the first day or two of a move that began with a price shock, allowing for the possibility of holding the wrong position in the future

and losses are critical to the net performance of any system; therefore, you should make a special effort to study those trades. In doing that, you may find that the largest profits were the results of price shocks, as discussed in the preceding point, or that there is an odd piece of data that was not obvious before. If there are a number of large profits and no large losses, you may have excellent risk control, but more likely you have overfit the data. There is an unfortunate tendency to look for problems when results are bad, but to accept them when they are good.

Reading between the Lines

It is easy to show specific examples of rules that produce large profits, while the overall strategy is bad. Many books base intuitive proof or justification on specific examples. This can work if the general theory is sound, when there is fundamental or economic substantiation, such as expecting a seasonal low in the U.S. crops at harvest. For patterns and many purely technical indicators, such as those using momentum, this is not as clear. Results are likely to show that the situations that fail greatly outnumber the fewer exceptionally good examples.

Systems That Work in Only One Market

From time to time, all traders receive mail offers for a highly specialized system called Cattle Trader," the "Silver Day Trading System," or "Easy Profits through Stock Index Trading." It is most likely that each of these systems has been finely tuned with rules unique to this one market. Once aware of the optimization process and its results, these offers must be viewed more critically; after all, with optimization techniques able to test combinations of indicators and rules, you could just as easily create a system that seemed to make as much on a single market. To prove that you have actually found something that works requires an understanding of the rules and the way it was tested, or time to watch the System operate in the real market. If possible, a comparison of the real trading profile compared with the expected performance based on testing should be available.

Student tTest

Among tests that help determine whether the results are significant, the student ttest is one of the most useful. It tells you whether there is a systematic bias in the data by showing whether the mean of the data is significantly different from zero. This is a useful piece of information when you try to decide whether the results of only a few trades represent a good system, or whether a series of losing trades implies that a system has no value. For a single set of trades produced by computer testing or by actual trading, the ttest is:

$$t = \frac{\text{average trade results}}{\text{1 standard deviation of trade results}} \times \sqrt{\text{number of trades}}$$

The critical value of t, which can be found in Appendix A, depends on the number of trades in the sample. For the case of a series of trades produced from the same market and same system, the degrees of freedom = number of trades - 1. Using the table in Appendix A if there were only 10

trades, the degrees of freedom would be 9, and the value of t must be greater than 1.833 to reach the 95% confidence level.

Where you want to know if two systems are producing results that are significantly different from each other, the technique is more complicated. It requires comparing the

~ Ben Warwick, Event Trading (Irwin, 1996).

525

mean, variance, and number of trades of the two systems; most important is the approximation for the degrees of freedom, which allows you to determine the confidence level of the results:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

where $\bar{x}_1$ = the average of method 1 trades

$\bar{x}_2$ = the average of method 2 trades

s_1 = variance of method 1 trades

s_2 = variance of method 2 trades

n_1 = number of trades in method 1

n_2 = number of trades in method 2

The calculation of degrees of freedom uses Satterthwaite's formula:

$$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{(n_1 - 2)} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{(n_2 - 2)}}$$

POINTANDFIGURE TESTING

The pointandfigure system is representative of the group of swing systems. Buy and sell signals occur when prices move above or below previous highs and lows, respectively. The current market direction changes when a price reversal exceeds a specified cumulative price change,

designated in points. Because it is a charting method, the reversal is shown in terms of boxes; the reversal criterion traditionally is expressed as a "3box reversal (for a complete explanation, see Chapter 11).

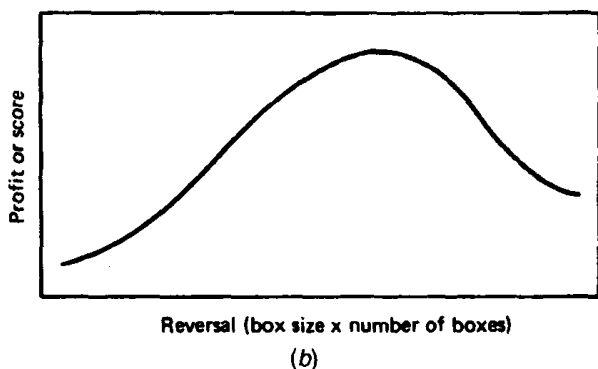
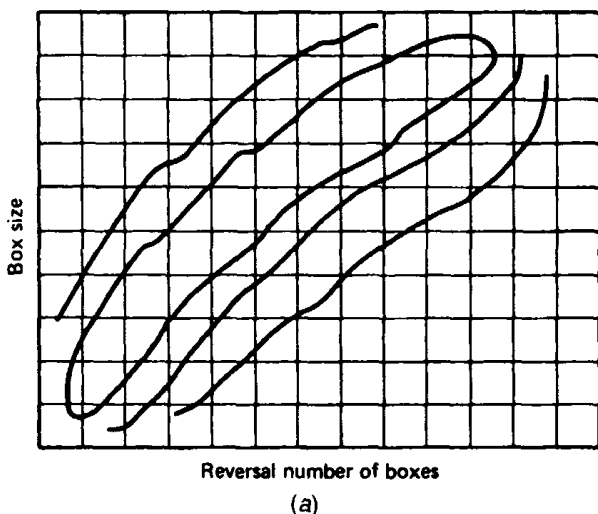
The size of the box and the number of boxes that define a reversal are the two variables that make up the reversal criteria. If silver were the object of the test, the box size might start as low as 10 points (1/10th of one cent, the minimum move) and increase in increments of 10 points; the number of boxes in a reversal would be the integer values 1,2,3. Figure 2 11 la shows that the expected pattern of test results is symmetric, extending across the diagonal drawn from the top left to the bottom right corners. This example followed the convention of putting the fastest trend change in the upper left (smallest box size and a 1box reversal) and the slowest in the lower right corners.

The symmetry is the result of very similar performance for equal reversal values, the product of the box size, and the number of boxes in a reversal. For example, a trend change of 40 points for the Swiss franc could have occurred using combinations of 40 points x 1 box, 20 points x 2 boxes, 10 points x 4 boxes, 4 points X 10 boxes, 2 points x 20 boxes, and 1 point x 40 boxes (among others). The difference in profitability occurs because the selection with the smallest box size will satisfy the penetration criteria sooner, giving slightly faster signals and more trades. The average profitability of all combinations of reversal size will appear as in Figure 2 11 lb. This representation can be used with the 2dimensional map to find the most consistent set of parameters.

When tested over long periods, most markets have increasingly wide price ranges, as clearly seen in the S&P Testing for fixed combinations of box size and number of reversal

526

FIGURE2111 Pointandfigure mapping strategy. (a) Reversal number of boxes. (b) Reversal (box size x number of boxes).



boxes for the past 10 years would find the single parameter set that performed best over fits and

this period; however, the increasing volatility would result in much larger profits and losses in recent years. This change of risk will force the optimized results to favor parameters that performed best in the last few years and might ignore the profits and losses of the earlier period, which would be relatively small. To put it in perspective, a 10% move in the S&P in 1987 would only be about 25 points, equal to a total of \$12,500, while it would be 90 points in 1997, or \$45,000. If prices moved 1% per day (3.00 points) in 1987, a reasonable reversal combination would be from .45 points for fast trading (.15point box with 3box reversal) to a full day's move of 3.00 (1.00point box with a 3box reversal). In 1997, either combination is likely to produce a new trade every few hours, creating a very different performance profile than prior years.

The problem of volatility was discussed in Chapter 11. The

alternative to a fixed box over time was to vary the box size as a percentage of price. Therefore, if the best reversal in 1987 was .45 for the S&P at a value of 300, it would be 1.35 in 1997 at 900. However, this may not have increased enough, and many market analysts treat volatility as an exponential function with respect to price. This can be approximated using a square root function or log function, both readily available in a spreadsheet or testing program:

527

$$\text{Box size} = p \times \sqrt{\text{price}}$$

$$\text{Box size} = p \times \ln (\text{price})$$

where p is the percentage that is varied as a parameter in the optimization. This last method would be similar to the technique of using pointandfigure on a log chart.

COMPARING THE RESULTS OF TWO SYSTEMS

The easiest way to understand the use of testing is by example. Using daily IMM Swiss franc data from January 1986 through February 1996, we compare the results of two trend systems, one based on a moving average and the other on exponential smoothing. These calculations can be found in Chapter 4 ("Trend Calculations"). In both cases, the system generates a buy signal or a sell signal when the trendline turns up or down, respectively. Table 211 shows the results of varying the calculation interval from 10 to 250 days in steps of 10 days. Parts a and b give the results using a commission of \$25 per trade, and parts c and d use \$ 100 per trade.

Which System Is Better?

If we scan the results of the net profit (NetPrft) in column two of parts a and b, we can find that the largest profit was \$52,125 for the exponential smoothing at 120 days, the largest return on account (ROA) was 376% for exponential smoothing at 160 days, and the highest profit factor was 3.47 for the moving average at 250 days. However, looking at individual values can be confusing and misleading. Choosing the one best performer is only sensible if you can guarantee that those parameters will be the ones that perform best in the next periodan impossible request.

At the bottom of each part of Table 2 11 are the average and standard deviation of each set of tests. For parts a and b, the average net profits for the moving average system were slightly better than exponential smoothing; however, looking at all the averages and variance of the two systems show that the results are nearly identical. From a practical point of view there is really no difference between the performance of the two systems. One might profit from one particular situation in one system, but over the long run they will be extremely close. This may be easier to see in Figure 2112, which shows the net profits of the two methods on the same chart. it may seem that the moving average technique performs slightly better with shorter intervals (30 to 100 days) and the exponential at longer ones (120 to 170 days), but these distinctions are very small. The overall picture is that a trending system produces profits based on the time interval

used in the calculation, and is not dependent upon the specific method of calculation. This observation allows us to view the robustness of trend following as a trading approach.

A comparison of optimization results usually selects the one with the highest profits, but that may not be the best choice. Comparisons should account for the following minimum criteria:

1. A good sample of parameter combinations were tested. This will include those variables that cause frequent, as well as slow, trading. The distribution of fast and slow test results should be similar, which may involve scaling the calculation periods in different ways. One simple way of knowing if you have succeeded in this distribution is by observing the average number of trades per test on each system.
2. The average results of all tests and the variance of results should be compared. Significantly higher and more consistent overall results is a strong argument for a fundamentally sound approach.

528

TABLE211 Comparison of Swiss Franc Trend Tests Swiss Franc
19861995, 10 years

(a) Moving Average with \$25 Commission				
Period	NetPrft	PFact	ROA	MaxDD
250	39063	3.47	354	-11038
240	3900	1.07	12	-31700
230	22350	1.48	129	-17263
220	25950	1.54	136	-19088
210	25900	1.46	87	-29725
200	20213	1.44	68	-29575
190	7388	1.16	20	-37113
180	28438	1.74	106	-26825
170	38150	2.26	283	-13463
160	34363	1.82	110	-31238
150	30475	1.62	150	-20375
140	37363	1.79	180	-20738
130	40988	1.88	200	-20488
120	39488	1.65	133	-29750
110	17725	1.24	62	-28463
100	44513	1.68	294	-15138
90	21088	1.25	73	-28988
80	40988	1.60	330	-12425
70	33050	1.44	171	-19313
60	34300	1.39	150	-22850
50	22488	1.23	79	-28413
40	46563	1.49	317	-14700
30	46175	1.40	269	-17163
20	16850	1.10	47	-35963
10	5763	1.03	23	-25150
Average	28941	1.57	151	-23478
Stdev	12213	0.48	101	7333

(b) Exponential Smoothing with \$25 Commission				
Period	NetPrft	PFact	ROA	MaxDD
250	37363	2.81	339	-11038
240	11013	1.26	38	-28763
230	26150	1.67	170	-15388
220	20963	1.41	99	-21163
210	27400	1.57	89	-30725
200	9300	1.17	23	-39988
190	18600	1.42	55	-33863
180	23088	1.56	91	-25238
170	45388	3.06	336	-13500
160	45863	2.42	376	-12188
150	39338	1.89	181	-21788
140	38313	1.82	174	-22063
130	36113	1.71	179	-20163
120	52125	1.95	272	-19138
110	12250	1.15	29	-42050
100	35800	1.48	254	-14088
90	30163	1.40	117	-25788
80	29875	1.37	199	-15000
70	7725	1.08	24	-31588
60	37538	1.44	124	-30363
50	26563	1.29	128	-20788
40	37538	1.37	193	-19400
30	30813	1.24	116	-26600
20	23938	1.15	81	-29638
10	-8975	0.96	-20	-44325
Average	27770	1.58	147	-24585
Stdev	13709	0.51	104	9056

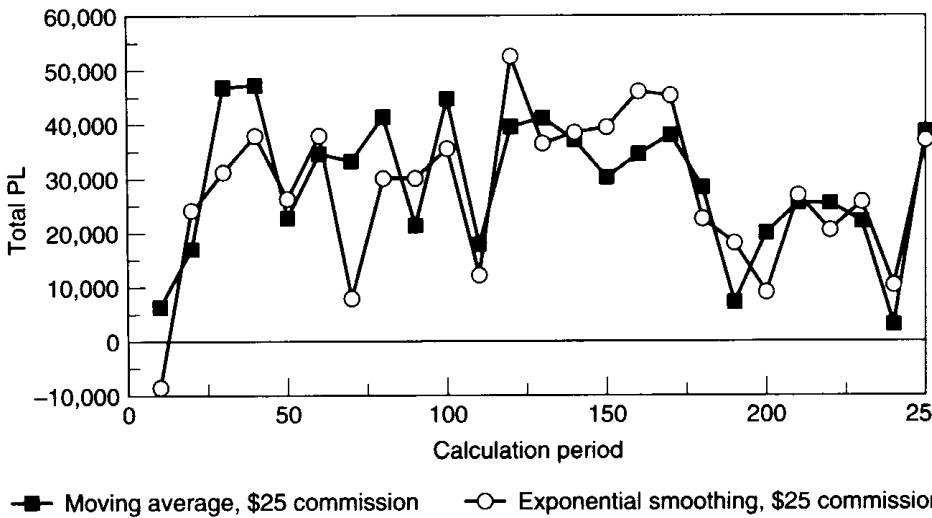
(c) Moving Average with \$100 Commission

Period	NetPrft	PFact	ROA	MaxDD	#Trds
250	36363	3.07	325	-11188	36
240	-675	0.99	-2	-35525	61
230	18825	1.38	92	-20488	47
220	21675	1.42	100	-21775	57
210	21025	1.36	66	-31900	65
200	16688	1.34	51	-32725	47
190	3563	1.07	9	-39888	51
180	25363	1.63	89	-28475	41
170	34325	2.04	227	-15150	51
160	29188	1.64	85	-34463	69
150	25300	1.48	110	-23000	69
140	32338	1.64	141	-22913	67
130	36188	1.72	161	-22438	64
120	33788	1.53	104	-32600	76
110	10975	1.14	35	-31163	90
100	36338	1.51	209	-17350	109
90	12613	1.14	40	-31763	113
80	33413	1.45	260	-12875	101
70	25175	1.31	112	-22463	105
60	25675	1.28	96	-26675	115
50	14163	1.14	48	-29763	111
40	36138	1.36	224	-16125	139
30	32900	1.27	176	-18738	177
20	-1000	1.00	-2	-42563	238
10	-19738	0.92	-57	-34575	340
Average	21624	1.43	108	-26263	98
Stddev	14127	0.42	89	8323	67

(d) Exponential Smoothing with \$100 Commission

Period	NetPrft	PFact	ROA	MaxDD	#Trds
250	34213	2.51	306	-11188	42
240	7188	1.16	23	-31538	51
230	22775	1.56	142	-15988	45
220	16388	1.30	71	-23088	61
210	23575	1.46	73	-32450	51
200	5625	1.10	13	-42763	49
190	14475	1.31	40	-36413	55
180	20013	1.46	74	-26888	41
170	42313	2.77	284	-14875	41
160	41288	2.17	286	-14450	61
150	34313	1.72	142	-24100	67
140	33438	1.66	138	-24313	65
130	30638	1.56	129	-23675	73
120	46125	1.79	219	-21025	80
110	4750	1.05	10	-45650	100
100	27325	1.34	162	-16863	113
90	21838	1.27	77	-28488	111
80	22150	1.25	125	-17700	103
70	-1650	0.98	-5	-36088	125
60	29513	1.32	88	-33663	107
50	17338	1.18	72	-24088	123
40	26213	1.24	126	-20825	151
30	16938	1.12	60	-28100	185
20	6388	1.04	18	-35250	234
10	-36125	0.85	-62	-58600	362
Average	20282	1.45	105	-27523	100
Stddev	16746	0.45	92	10798	71

Comparison of Swiss franc tests 1986–1995



3. Results should be scored in some way to include profitability and risk. Other measurements can be used, but stability is of great importance.
4. Results, or scores, should be averaged to avoid spikes or unusual distortions.
5. Execution costs, including liquidity factors, must be included in the manner appropriate to each system. A trendfollowing system with entries and exits in the direction of the trend should have greater slippage than a countertrend entry method.
6. Make mistakes in the direction of being conservative. If two systems have comparable returns, select the one with the lower risk. If they have similar returns and risk, choose the one with fewer trades.

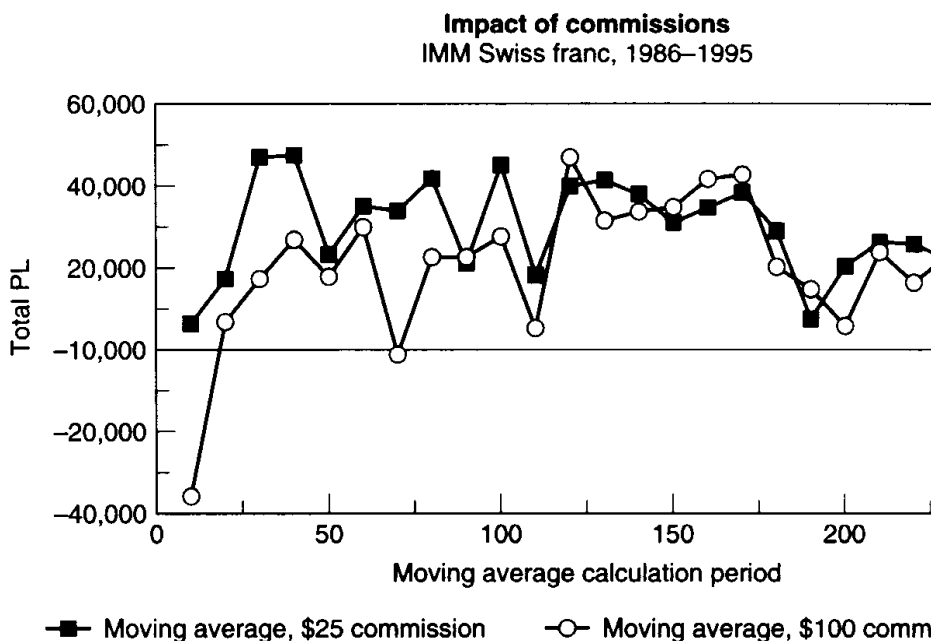
Choosing the speed of the trading system is often the most critical decision in the performance profile. Parameters that cause a trend method to trade faster are most desirable, because the higher number of trades adds confidence to the validity of the test results. You might also find that there is more consistency in the performance of a faster trading program compared with a slower one that generates the same net profits. To compensate for its advantages, faster trading is more susceptible to execution problems, so that performance will deteriorate quickly if slippage increases. Figure 2113 shows the change in performance of all test intervals from 10 to 250 days for the moving average method when transaction costs are increased from \$25 to \$100 per trade. Note that the shortest interval turns from a small profit to a large loss and profits for intervals through 40 days are halved. Some areas of particularly poor performance, such as at 70 and 110 days, are magnified by the increased costs. In general, there is very little impact on the performance of slower parameter selections.

PROFITING FROM THE WORST RESULTS

Under the right conditions, the worst results of an optimization or test sequence may hold valuable information; it presents the worstcase scenario that should be carefully reviewed for system problems, in particular with regard to risk control. If the area of poor results is broad and continuous, it may also show a way to improve entry and exit timing. Figure 2 114a and b shows the continuous results of 1 and 2variable tests. Both identify faster trading areas as having maximum loss and slower trends as having best performance.

531

FIGURE 2113 Impact of increased transaction costs on performance.



The worst results are only useful if there is a net profit from taking the opposite position. This is the case only if the net loss is greater than the transaction costs, including commissions and slippage. But even if the worst score has a nearzero return before transaction costs, it can be used profitably. Unprofitable results of a fast trend mean that a long or short position taken at that point lasts only a few days and does not indicate a sustained trend. Use the following rules:

1. Trade a system of two moving averages, a longterm, designated by the best score D2, and the shortterm, selected as the worst score D1.
2. When prices cross the longterm moving average, a long position becomes eligible. A long position is entered when the shortterm moving average signals a short signal. Positions may be entered on a 1day delay.
3. Short positions may be closed out when prices cross above the longterm trendline; they must be closed out when a long position is entered.
4. Short positions and exits from long positions are the opposite of rules 2

and 3.

Because the shortterm signals are not good indicators of trends, they can be used as a countertrend entries. That allows for better execution when trading Larger positions or in less liquid markets. The choice of immediate exit or delayed exit is the trader's preference. Once a position is held, there is greater risk waiting to exit, yet most exits would be improved by better timing.

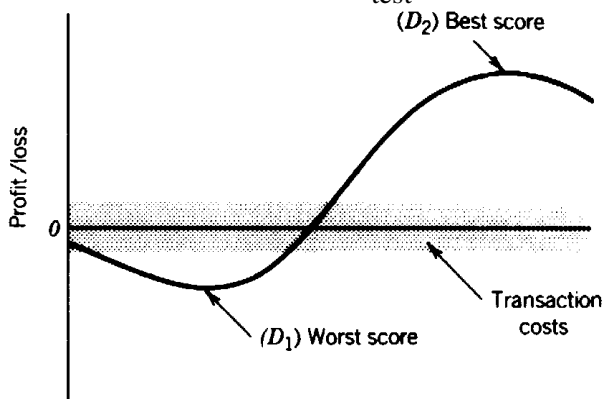
RETESTING PROCEDURE

One optimization is never the end of the testing process. If you have successfully finished the development of a system, and have retained the most recent data for outofsample verification, then you should retest the system including the outofsample data before beginning to trade the program. Any additional data adds robustness to the system. The impact of adding data will depend on the nature of the market during the new period and the amount of data originally used in the testing.

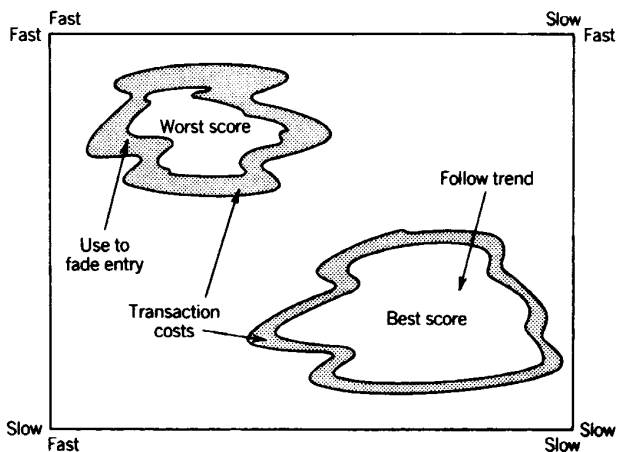
As in stepforward testing, some analysts prefer to test a fixed amount of data, dropping off the oldest. This does not mean that they test only 6 months, or 1 year of data and

532

FIGURE2114 Dual use of test map. (a) Singlevariable test. (b)Twovariable test



(a)

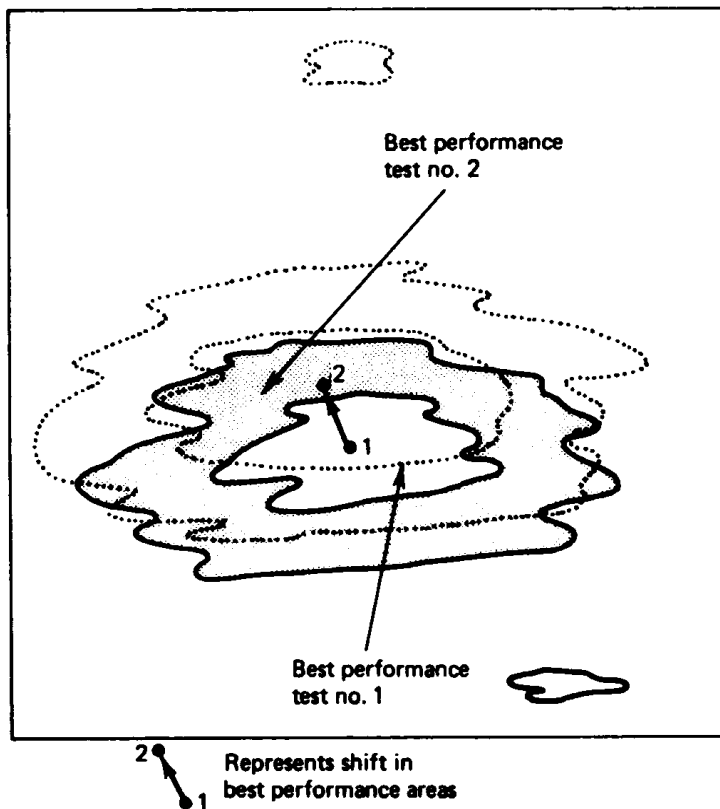


(b)

retest every month. You may decide that data before a particular date is no longer relevant. In some cases, there are structural changes that support that approach, such as the formation of a unified European currency, which would put controls on the variance in exchange rates between the participating countries even before the inception date. You might also consider a new market with low liquidity during its earlier years as different from more active, current market conditions. Retesting remains an important way to adjust the system to characteristics of new data. You might decide that retesting is necessary when time representing 5% to 10% of the test period has elapsed. Using the 2variable test as an example, a shift may be expected in the area of best performance as shown in Figure 2115.

Because only a small amount of data was added, the size of the shift in Figure 2115 should be small. If a large shift actually occurs, then two possibilities should be considered:

1. The data was unusually volatile and introduced patterns not previously seen in the data. In this case, the shift in parameters is justified.
2. The data period for the original tests was short and did not include enough patterns to make the parameter choice robust. With a small amount of test data, there is a



greater likelihood that a faster trading model will be selected. As the data shifts and that faster model is no longer good, the parameters may jump to a very slow model. This erratic pattern of shifting from fast to slow is indicative of too little test data.

Optimizing after a Major Move

When your system fails to produce the right amount of profit from a major move, or when the stock market takes an uncomfortably large plunge, it is natural to question whether the system has the right parameters. Is the trend fast enough? Is the stoploss too close? This is the same process as retesting. If you have used enough data for your original tests, then it is very likely that the new tests won't change anything. If you make up new rules to fix this one problem, you are overfitting and accomplish nothing. If you have doubts about the test period, use more data rather than less. The new patterns will only give you a more realistic idea of the market risk than you had before. A major move that contains a price shock is a special situation discussed later in this chapter, as well as Chapters 22 and 23.

COMPREHENSIVE STUDIES⁶

A comprehensive study points out the robustness of a trading strategy. When a system performs well in many markets using similar

calculation intervals, it is fair to assume that the method is sound. If a technique works in the Swiss franc but not the Deutschmark, or in Euroyen but not Eurodollars, you must be able to understand the significant differences

6 Readers with an interest in performing optimizations such as these should read the basic material in Chapter 4 and the additional systems in Chapter 5.

534

before accepting the method. Markets have individual characteristics that might justify differences in performance; however, a strategy that can generate profits in a broad range of markets is robust. In addition to the studies that follow, Chapter 5 ("Trend Systems") includes a section "Comparison of Major Trend Systems," which will be helpful to anyone interested in viewing the results of different systems side by side.

Colby and Meyers, in *The Encyclopedia of TecAmical Market Indicators* (Dow Jones Irwin, 1988), have given a very useful comparison of major market indicators simply by presenting all of their test results in a common form. Any one of the methods may have proved good or bad during the past 10 or 20 years, but taken as a whole, the consistency of various techniques adds to the robustness of that approach. For example, if one volumebased indicator was profitable, while four others generated consistent losses, you would need to understand why that one method worked before it could be used; the evidence of the four poor indicators argues that the use of volume in this way is not robust.

There are three previous, wellknown comprehensive studies of moving average systems by Maxwell, Davis and Thiel, and Hochheimer. In addition, optimizations of Wilder's Directional Movement and RSI will be found in the sections that discuss those techniques. Although dated, these studies are very similar in structure to the way anyone would begin to test a trading idea. they emphasize different features of trading that were important to the authors and give us useful insight. By virtue of hindsight, we can see that along with completeness in one area often comes a deficiency in another. Davis and Thiel analyze the greatest variety of markets, covering virtually all of the U.S. crops as well as cattle, eggs, and the soybean complex. They include about 5 years of data and use relatively fast moving averages (up to 10 days); they introduce variations in leadoriented plotting and in testing nonconsecutive days; the data used are only the closing prices. The results are clearly presented in both detail and summary, and yield generally good returns.

The authors R.E. Davis and C.C. Thiel, Jr., present excellent credentials and experience in systems testing. Their study of moving averages uses combinations of simple buy and sell signals, leading plots, and skips in the selection of sequential prices used (e.g., a skip of two uses every other price). They test a total of 100 combinations of the three factors:

A skip from 1 day (none) to 5 days (1 week)

An average of 5, 6, 7, 8, or 10 days

A leap of 0 or 2 days

Markets tested were soybeans, bellies, cattle, cocoa, copper, corn, eggs, soybean meal, oats, soybean oil, potatoes, rye, sugar, and wheat.

Maxwell's study is extremely comprehensive but limited by its application to only pork bellies. His idea was to apply combinations of features and test the results. The first feature was the choice of trend and included the possibilities of a simple mean, or a moving average of 3, 5, or 10 days, of either a conventional, averagedmodified, or weighted type. The second feature was the delay factor, used to improve timing of both entry or exit. Some of the possibilities were: (1) act without delay; (2) act if the signal condition persists for one additional day, (3) enter if the signal condition persists for 2 additional days, but liquidate without delay; and others. Combinations of two types of moving averages and a delay factor were tested, with and without fixed or moving stops. With 10 types of averages, 6 delay factors, and different stops, Maxwell has a lot of combinations to examine.

The study is then expanded to 3factor systems with a list of 18 combinations of rules to generate 324 systems, of which the results of 285 are recorded, with 49% profitable and 51% losers. The largest loss was generated by a system with the rules:

1. Enter a new position when both the weighted 3day and weighted 10day moving averages cross the pricemean, as long as the signal condition persists for 2 additional days (buy if averages cross moving down, sell if up).

535

2. Liquidate positions when the 3day average reverses its direction through the pricemean as long as it continues for 1 additional day.
3. No fixed stops are used.

The best profits in the 3 factor system required both a 5day averagedmodified and a 10day weighted average to move across the pricemean, provided the 5day average lagged behind the 10day. The current position was liquidated when the shorterterm average crossed the longer term. No fixed stops were used.

Maxwell's study represents a great amount of work and some simple and sound philosophy, but does not cover an adequate sampling of data to justify most of his conclusions. Conclusions were drawn based on a selected 50day test period using May 72 pork bellies, leaving many questions regarding the success of this system or any other system when applied to this short test interval. Maxwell does test four other selected 50day periods, but the reader cannot know how these periods were chosen. Considering the effort in outlining a testing program and establishing rules, the extent of the system testing is disappointing.

Comparing Methods of Calculation

Hochheimer:⁷ performs two interesting studies: a comparison of three types of moving average calculations and a test of the use of channels and crossovers. In the first analysis, Hochheimer compares simple, exponentially smoothed, and linearly weighted moving averages, tested on a good sampling of markets (without financials and currencies, which did not exist during the 1970 through 1976 test period). The simple moving average and exponentially smoothed calculations are covered in Chapter 4 ("Trend Calculations"), and the linearly weighted refers to stepweighting,

in which integer values are used and each successive day is incremented by one, the most recent day having the highest weighting factor.

The rules that apply regardless of which calculation was performed, were:

1. The trend calculation used the closing price only.
2. A buy signal occurred on an upward penetration of the moving average by the closing price; a sell signal occurred under the opposite conditions.
3. The model is always in the market.
4. A trade cannot be executed when the day's high and low prices are the same. It is assumed to be a locked limit day.

The test covered 7 years of data from 1970 through 1976 and moving average days (or equivalent smoothing constant converted as $2/(\text{days} + 1)$), which appear to range from 2 to 70. The results shown in Table 212 are interesting and logical, considering the rules. The longer trends (using more days) were consistently better than shorter ones, regardless of which technique was used. This would imply that the longer trends are more reliable. With a few exceptions, the best selection of days ranged from 40 to 70.

A more careful look at the results proves interesting. During the 7 years of test data, soybeans had 728 trades using a 55-day simple moving average. That is more than 100 trades each year, one every 2 or 3 days. Hogs had 1,093 trades during the same period, and other commodities were similarly high. That seems to be an unusually large number of trades for the time interval used, indicating that there must have been many false trends that resulted in small losses. This is confirmed by the very low percentage of profitable trades, only 21% for soybeans. Nothing is indicated about transaction costs, and 728 trades costing \$50 per trade in commissions and slippage would generate \$36,400 in costs, about 15% of the Net

Frank Hochheimer, "Moving Averages," and "Channels and Crossovers," in *Technical Analysis in Commodities* Perry J. Kauffman (ed.) (John Wiley & Sons, New York, NY 1980).

	Best Day	Best Range	Net P/L	Max. Run of Losses	Total
Cocoa	54	53-59	\$87,957	\$-14,155	600
Corn	43	43-46	24,646	-6,537	561
Sugar	60	55-60	270,402	-15,563	497
Cotton	57	52-57	68,685	-11,330	64
Silver	19	—	42,920	-15,285	1,397
Copper	59	54-59	165,143	-7,687	437
Soybeans	55	55-60	222,195	-10,800	720
Soy meal	68	65-70	22,506	-20,900	704
Wheat	41	40-45	65,806	-12,550	480
Pork bellies	19	16-25	97,925	-9,498	774
Soy oil	69	65-70	89,920	-8,920	580
Plywood	68	65-70	1,622	-3,929	377
Hogs	16	16-20	35,595	-7,190	1,097

P/L shown in the table. The total soybean profit of \$222,195 was equal to \$305 per trade profit over the test period, a very good net return and one that might have been unusual because of the exceptional volatility of the agricultural markets during the early 1970s.

Comparing the Hochheimer results with the same system applied to the 10 years from 1986 to 1996, we see a very different picture (see Table 213). Many of the markets that produced large profits now show losses. The currencies stand out as uniformly successful; however, the time periods for calculations vary from 40 to 75 days. The appearance of large calculation periods for the "Best Day" column indicates either a very successful long trend, as in the T-bonds, or an attempt to minimize the losses of markets without profitable trends.

In most cases, you will find that the shorter trend periods are unstable. Because consistent shortterm trends are rare over long test intervals, such as 10 years, profitable results for a 15-day moving average are usually caused by one or two very large profits, or a sustained move in one direction. The choice of a shortterm trend rarely produces profits in the future.

The ease of new computerized testing platforms allows a convenient look at robustness. For the single moving average, 20 periods from 5 to 100 days were tested in steps of 5 days. The last column of Table 213 gives the number of profitable tests. Six of the 15 markets tested had no profitable results using any calculation period. Overall, 45% of all tests were profitable. The interest rates show very erratic results with a wide range of best days, even for the T-notes and T-bonds, which had a large number of profitable combinations. Only the currencies appear to be robust within a framework of moderate performance.

Channels and Crossovers

This Hochheimer study consisted of two tests. The first is a crossover of two simple moving averages. The shortterm average (fast trend) ranged from 3 to 25 days and the longterm from 5 to 50 days. The objective was to eliminate the whipsaws that were evident in the first study. The rules for this system were:

1. Each moving average used closing prices only; the longterm number of days was always greater than the shortterm days.

537

TABLE 21-3 Results of a Simple Moving Average Model (1986–1996)

<i>Market</i>	<i>Best Day</i>	<i>Net P/L</i>	<i>Max Loss</i>	<i>Total</i>	<i>Trades</i>
Corn	80	-937	-8,462	98	23
Cotton	60	62,060	-76,640	99	38
Soybeans	95	-1,825	-11,175	124	33
Silver*	40	-485	-12,720	112	24
Copper*	45	18,987	-9,000	89	34
Gold	45	-3,240	-16,500	146	44
Swiss franc	45	76,025	-15,000	125	54
German mark	75	46,462	-13,462	91	29
Japanese yen	60	85,475	-13,737	169	61
British pound	40	61,825	-33,237	158	54
S&P 500	35	-4,650	-75,750	200	68
NYSE Index*	90	-7,100	-21,350	88	29
T-Bills	30	8,350	-6,325	156	49
T-Notes	15	52,215	-9,235	267	102
T-Bonds	100	51,084	-14,036	74	33

*Test data began in 1987 or 1988.

2. A buy signal occurred when the shortterm average moved above the longterm average; a sell signal occurred when the shortterm average moved below the longterm average.
3. The model was always in the market.
4. if the high and low prices of the day were equal, a lockedlimit day was assumed and no execution occurred.

The use of a longer and shorter trend introduces the ability to identify a trend bias in the market and trade only in that direction. Theoretically, this should apply to the interest rates and to the index markets, which have either sustained trends or an upward bias. During the period of the original tests, it may be the seasonality of the agricultural markets that dominate the trend. Using two moving averages, the faster one can be considered a timing mechanism, while the slower one tracks the underlying market direction. In Table 214, the number of trades are less than the simple moving average model, and the percentage of profitable trades was increased, the type of results that indicate a more robust method.

Retesting

The testing performed to find the best set of parameters is never the last test. it is always important to monitor performance and compare the outofsample results with a new optimization to determine the success of the system rules and the testing process. A 1982 publication of Hochheimer" provides a rare opportunity to see how the original tests fared and how parameter selection would have changed based on both additional data and changes in price patterns. We can do this ourselves by testing 5 or 10year periods and comparing the results of the same portfolio. From these results we can visualize an overall change in

Frank L. Hochheimer and Richard J. Vaughn, Computerized Trading Techniques 1982 (Merrill Lynch Commodities, New York, 1982).

538

TABLE214 Results of Two Simple Moving Average Crossovers

	Best Days	Net P/L	Max. Run of Losses	Total
Cocoa	7 25	\$176,940	\$-4,436	468
Corn	11 47	69,275	-2,697	225
Sugar	5 50	335,843	-13,851	311
Cotton	16 25	304,485	-4,755	485
Silver	3 26	100,790	-8,610	661
Copper	17 33	212,939	-3,057	354
Soybeans	16 50	286,440	-15,665	311
Soybean meal	18 50	117,155	-8,155	272
Wheat	11 47	113,118	-2,660	209
Pork bellies	25 46	13,124	-21,538	226
Soybean oil	14 50	121,749	-6,585	327
Plywood	24 42	18,505	-3,184	219
Hogs	3 14	97,448	-7,805	793

which the market moves. We may find that noise has increased, causing a shift to longer trends, or that risk is much greater and the returns are proportionally smaller.

Crossovers Retested

In the 1982 update of the Crossover System optimization using data through 1979, it is interesting to note that 11 out of 17 markets selected two moving average speeds either identical (in six cases) or nearly identical to the previous test. Five of the remaining six tests showed new selections that traded slower than the previous test. Only gold resulted in a faster selection, but the original test in 1976 could have had only one year of data.

Because the results of using the parameters selected in 1976 were not shown in the new study, their performance could not be assessed. Eleven cases, however, in which the new parameters were either identical or had very small changes, can be used as a conservative estimate. Their outofsample performance is shown in Table 215. If the parameters are not

identical, the 1979 test must show results that are higher than the actual outof sample would have been.

The first reaction to the update of the 1976 crossover test should be very positive. Almost identical parameters for 11 out of 17 products originally tested were selected. This means that the traders who used this system were using the optimum parameters during the 3year period from 1977 through 1979. How did they do?

Table 216 is a comparison of the outofsample period, 1977 through 1979. In the cases in which the parameter selection was slightly different, it must be assumed that the total performance, 1970 through 1979, was better than the performance using the original parameters. Therefore, the outofsample period should be at least as good as the results actually achieved. Total profits were \$201,649, trading one contract of each market.

Two additional factors should be considered. Although only 3 of the 11 showed losses, the fast nature of this system results in very low profits per trade in 4 of the remaining 8 markets. This may not seem serious until the realities of trading are considered, where the exact execution price is rarely achieved. A trendfollowing system, with orders entered as intraday stops, must always allow for slippage caused by lack of liquidity and directly related to the size of the order being executed (see the discussion of liquidity in Chapter 18). In a fast market, this slippage can be large, often hundreds of dollars (in contract

539
TABLE215 Retesting of the Crossover System
Through 1976

	Best Days	Net P/L	Trades Profitable/ Total	%	Profits/ Trade	Best Days	Th
Corn	11 47	\$ 69,275	93/225	41	\$ 308	12 48	\$
Sugar	5 50	335,843	109/311	35	1080	6 50	3
Cotton	16 25	304,485	223/485	46	628	16 25	3
Silver	3 26	100,790	262/661	40	152	3 26	
Copper	17 33	212,939	177/354	50	602	17 32	2
Wheat	11 47	133,118	87/209	42	637	12 48	1
Pork bellies	25 46	13,124	100/226	44	58	25 46	
Soy oil	14 50	121,749	128/327	39	372	14 50	1
Plywood	24 42	18,505	100/219	46	84	20 46	
Cattle	7 13	147,540	337/792	43	186	7 13	1
T-Bills	6 18	39,710	69/148	47	268	6 18	

value) away from the intended price. Even MarketonClose orders are subject to slippage. A buyer using a CloseOnly order should expect to get the high of the closing range; a seller will get the low. in Table 216, the furthest right column assumes a conservative slippage, one that should be included in any test program. In some cases, such as copper and bellies, the slippage is small and does not seem to impact on profits; even the large

profits of cotton survive well. But in the fastest trading commodities, even the smallest slippage will severely reduce large profits, frequently turning them into losses. The final total shows that those small costs, when applied to every product, were greater than the profits. Instead of making a killing in the market, these traders netted a loss.

TABLE 21-6 Out-of-Sample Results for the Crossover System

	Net	Trades Profitable/ Total	%	Profits/ Trade		
	P/L					
Corn	\$ 14,311	36/107	34	\$ 133*	1/2¢	=
Sugar	12,990	51/138	37	94*	10 pts	=
Cotton	73,995	106/212	50	349	10 pts	=
Silver	-3,795	123/437	28	-9	1/2¢	=
Copper	5,851	68/217	31	27*	20 pts	=
Wheat	18,753	43/96	45	195*	1/2¢	=
Pork bellies	65,083	37/86	43	757	10¢	=
Soy oil	4,099	62/161	39	25	5¢	=
Plywood	-15,838	32/132	24	-120*	20 pts	=
Cattle	14,740	153/394	39	37	10 pts	=
T-Bills	-11,500	79/240	33	-48	2 pts	=
Totals	\$201,649					

*Best case (could have been no better, possibly worse).

540

Crossover System, 1986 to 1996

Bringing this technique forward to current markets, seen in Table 217, there is again a tendency to select longer calculation periods, especially for the shortterm trend; therefore, there has been a shift from the optimum trend speeds of 5 to 10 years earlier. The number of total trades has dropped significantly across all markets, and may be attributed to more market noise and less definitive trends. Because the use of two trends greatly increases the chance of overfitting the data, it should not be surprising that all of the markets show profits for some combination of fast and slow moving averages. The last column, indicating the number of profitable combinations, is needed to get a picture of the robustness.

Taking a closer look at Table 217, it appears at first that the S&P would be highly profitable; however, the profit of \$194,750, drawdown of \$63,650, and 63% profitable trades is deceiving because only 8 of 120 combinations were profitable. As a group, the currencies and interest rates performed well, and the total of all profitable combinations was 56%, showing that the technique of using two moving averages is likely to be more robust than the 45% given by the simple moving average. When making these system comparisons, you must be cautious that these tests

represent a comparable range of test speeds.

Channels

The next model relies on trends but are very different from the moving averages used in the previous optimizations. Originally called a price channel, it is now more familiar as an Nday breakout, the penetration of the high and low of the past M days. The penetration, which causes a buy or sell signal, occurs if the closing price penetrates the highlow band.

Table 218 shows the results of the interday test. A common variation of this method is to buy when the intraday high penetrates the upper band; however, this can cause ambiguities when both the high and low of the day exceed both bands. It is not possible to know which one occurred first. Intuitively, the closing price is expected to be a more reliable indicator of direction than the intraday high or low.

TABLE217 Results of a TwoMoving Average Crossover Model
(19861996), \$50 Transaction Costs*

<i>Market</i>	<i>Best Day</i>	<i>Net P/L</i>	<i>Max Loss</i>	<i>Total</i>
Corn	17 × 60	3,775	-5,375	44
Cotton	11 × 60	66,725	-5,390	40
Soybeans	20 × 25	73,450	-44,100	131
Silver [†]	15 × 19	18,230	-10,180	155
Copper [†]	15 × 25	17,337	-15,087	81
Gold	11 × 25	27,990	-8,900	93
Swiss franc	7 × 55	79,600	-13,912	54
German mark	25 × 50	52,387	-10,750	44
Japanese yen	11 × 20	121,362	-14,112	113
British pound	25 × 45	61,612	-19,025	62
S&P 500	20 × 23	194,750	-63,650	199
NYSE Index [†]	20 × 23	58,400	-31,475	179
T-Bills	19 × 45	12,100	-4,975	55
T-Notes	11 × 30	58,643	-7,368	79
T-Bonds	11 × 35	45,305	-14,513	67

*Fast trend tested from 3 to 25 in steps of 2; slow trend tested from 15 to 60 in steps of 5.

[†]Test data began in 1987 or 1988.

TABLE 218 Results of Interday Closing Price Channel

	<i>Best Day</i>	<i>Net P/L</i>	<i>Max. Run of Losses</i>
Cocoa	53	\$147,913	\$ -6,248
Corn	49	39,533	-5,048
Sugar	40	296,027	-20,758
Cotton	62	206,575	-6,870
Silver	14	27,690	-12,365
Copper	27	151,671	-7,225
Soybeans	51	244,839	-11,325
Soybean meal	49	104,690	-7,000
Wheat	23	111,087	-6,900
Pork bellies	52	60,263	-9,892
Soybean oil	51	8,646	-3,622
Plywood	49	8,632	-4,432
Hogs	9	83,702	-9,854

In the original tests shown in Table 218, the channel method has much lower profits and proportionally greater risk than the crossovers; results tend to be more erratic as well. This method is clearly better than the single moving average, with more consistent profits, fewer trades, greater reliability, and even smaller drawdowns.

Testing the breakout strategy for 1986 to 1996 (Table 219), we see a different pattern than the update of the moving average method. The number of trades have dropped, the maximum drawdown has increased, while only one market showed all

TABLE 219 Results of an NDay Breakout (1986-1996), \$50 Transaction Costs*

<i>Market</i>	<i>Best Day</i>	<i>Net P/L</i>	<i>Max Loss</i>	<i>Total</i>
Corn	55	5,275	-3,375	23
Cotton	35	54,605	-10,700	34
Soybeans	95	13,225	-6,150	12
Silver [†]	85	-2,685	-13,250	13
Copper [†]	40	9,400	-6,887	23
Gold	95	1,310	-14,430	14
Swiss franc	40	67,862	-16,600	31
German mark	40	42,775	-9,137	31
Japanese yen	25	76,062	-17,062	44
British pound	35	62,012	-13,625	36
S&P 500	85	15,025	-36,750	16
NYSE Index [†]	55	950	-21,200	17
T-Bills	30	10,875	-3,825	41
T-Notes	25	42,600	-9,087	53
T-Bonds	70	41,302	-20,323	19

*Breakout tested from 5 to 100 days in steps of 5.

[†]Test data began in 1987 or 1988.

542

losses. Of the 300 total tests, 57% were profitable, greater than either the moving average or crossover methods.

Modified ThreeCrossover Model

The use of one or more slowmoving averages may result in a buy or sell signal at a time when the prices are actually moving opposite to the position that is about to be entered. This may happen when:

1. The rules consider the crossover of the moving average, rather than a penetration of the price.
2. The change in a simple moving average value based on the new price is less than the change in the oldest price that was dropped.

For example, if a long signal occurs and the oldest price showed a decline of 50 points while today's price declined 40 points, the new moving average value will rise by the difference, +10, divided by the number of days in the moving average. This may also occur using exponential smoothing under the special conditions

$$E_{t-1} - P_t > P_{t-1} - E_{t-1}$$

By using a third, fastermoving average, the slope can be used as a confirmation of direction to avoid entry into a trade that is going the wrong

way. This filter can be added to any moving average or multiple moving average system with the following rule:

Do not enter a new long (or short) position unless the slope of the confirming

moving average (the change in the moving average value from the prior day to

today) was up (or down).

The speed of this third, confirming moving average only makes sense if it is equal to or faster than the faster of the trends used in the Crossover System.

Test Results

Fortunately, the results of both the Crossover System as well as the Modified ThreeCrossover System were available for the same commodities and the same years. Because the crossovers used in the latter model are exactly those of the first system, the comparison will show whether the confirming feature improved overall results. From the 22 markets tested, plywood was removed because its poor results on both systems tended to distort the comparison. The important statistics are shown in Table 2110.

The difference in using the Modified ThreeCrossover System versus the simpler

Crossover System are:

Average change in profits

Average change in equity drop

Average change in percentage of profitable trades

Average change in number of trades

A decline in profits equal to a decline in equity drop (or risk) is the same as using the original Crossover System with a 15.4% smaller investment. The percentage of profitable trades increased negligibly, indicating that the confirmation filter did not eliminate more losing trades as was expected. The last point, however, shows a larger decline in the total number of trades, indicating that the profit per trade has increased. The new filter appears to catch the entries at a better point, and eliminate some trades with smaller profits. This type of improvement means that there is more latitude for trading error,

	Crossovers				Modified Three-Crossover		
	P/L	No. Trades	% Drop	% Profitable Trades	P/L	No. Trades	% Drop
Deutschemmark	\$ 96,510	259	7.6	54	\$ 97,698	226	3.4
Japanese yen	94,575	131	3.1	51	92,287	111	2.9
Canadian dollar	71,940	158	7.0	52	69,690	116	5.2
British pound	80,262	113	8.2	46	69,808	81	8.3
Swiss franc	120,674	121	4.8	48	100,891	91	4.6
Cocoa	408,262	353	2.4	46	282,453	291	7.6
Coffee	83,586	332	3.9	39	56,338	231	5.8
Sugar	348,833	449	3.8	36	331,526	322	4.1
Cotton	378,440	697	2.8	47	282,460	483	3.3
Silver	96,995	1098	13.9	35	89,165	990	15.7
Copper	218,790	571	2.7	43	217,689	408	2.9
Soybeans	386,137	483	4.8	47	308,233	317	5.7
Soybean meal	165,294	429	5.3	43	135,643	316	5.9
Wheat	151,871	305	2.7	43	90,093	207	6.9
Pork bellies	78,207	312	27.5	44	76,363	246	14.2
Soybean oil	125,848	488	5.2	39	89,834	346	9.3
Hogs	88,097	853	9.2	39	94,634	647	7.9
Cattle	162,280	1186	4.8	41	155,135	875	3.3
GNMAs	76,291	197	4.4	52	56,217	133	6.5
T-bills	28,210	388	27.0	38	20,436	284	3.4
Gold	131,325	275	1.6	53	98,030	158	2.2
Average			7.3	45			6.1

such as slippage. Although the overall profile of the Modified ThreeCrossover is not much better than the Crossover System, it would be the preferred choice based on this information.

4918 Crossover Model Results

This model, using the same rules as the Modified ThreeCrossover System, was well known before Hochheimer's studies. It can be assumed that the selection of 4, 9, and 18 days was a conscious effort to be slightly ahead of the 5, 10, and 20 days frequently used in moving average systems. it is likely that the system was not developed by extensive testing, since all markets are traded with these same speeds.

As simple as it seems, there are a few very sound concepts in this approach:

1. Each moving average is twice the speed of the prior, enhancing their independence in recognizing different trends.
2. They are slightly faster than the conventional 5, 10, and 20day moving averages.

3. They are the same for all products, implying an attempt at robustness.

It would not be possible for either the profits or the equity drops to be better in this model than an optimized strategy, such as the Modified ThreeCrossover Method; yet, out of 17 commodity markets simulated during the 1970 through 1976 period, only two lost money. That is a very impressive record. There should be a greater degree of confidence in this performance than in an optimized program.

544

Importance of Failed Tests

When the period from 1986 to 1996 is tested, as seen in Table 2 1 11, we see much lower profits, large losses in the index markets, and uniformly high risk. Net results are worse than other methods tested. This certainly means that this technique is no longer good. Perhaps it means that a rigid set of trends cannot survive over the long term; the lengths of the underlying trends have changed; they are no longer uniform because of changing market participation; or there is more noise in the market, making it difficult to identify the trend in a timely fashion.

These results should make us rethink the conclusions drawn from the previous tests. We have looked at the best days and the robustness of the technique. Because we are no longer confident that the best day of the past will be the best day of the future, we need to depend on the robustness of the system even more. If a large percentage of all tests are profitable, then there is a much greater chance that any parameter selected will be profitable. Our goal is to find the most robust system.

To remove the dependence upon specific fixed parameters that might return unstable performance over many years, it is necessary to generalize many of the features in a system. For example, instead of a fixed stoploss at 10 points or \$500, it will be necessary to allow that stop to vary based on market volatility. Even the calculation period of the trend may change according to volatility, market noise, volume, or some other factors. This approach is discussed further in Chapter 17 ("Adaptive Techniques").

Test Criteria

The method used in these tests is not bad, but real market factors must be included or the results are deceiving. Although the researchers may have been pleased with the outofsample profits, they were only paper profits. In real trading, the chances greatly favor losses.

Testing should closely approximate trading. If there is doubt about the costs, make them larger. A system that can profit under testing penalties should make money when actually traded. Even when using prepackaged computer optimization software, the slippage

TABLE 2111 Results of a 4918 Crossover (1986-1996), \$50 Transaction Costs

<i>Market</i>	<i>Net P/L</i>	<i>Max Loss</i>	<i>Total</i>
Corn	-425	-12,087	114
Cotton	28,110	-21,000	119
Soybeans	-19,087	-33,987	139
Silver*	-20,740	-21,640	94
Copper*	6,862	-10,062	86
Gold	2,250	-12,910	130
Swiss franc	3,325	-43,350	124
German mark	11,562	-31,525	119
Japanese yen	86,412	-17,075	104
British pound	45,512	27,362	122
S&P 500	-76,650	-110,350	136
NYSE Index*	-31,475	-51,075	98
T-Bills	-7,025	14,400	125
T-Notes	15,674	-11,406	116
T-Bonds	-14,668	-44,013	130

*Test data began in 1987 or 1988.

545

can be added by making the commission costs high; both are per trade costs. If a mistake is to be made, do it by being too conservative; however, it is always best to be realistic.

Getting the Big Picture through Testing

As mentioned throughout this chapter, the key to robustness is to test more data, more markets, more of everything, and then to observe the results looking for a common pattern of overall consistency. There is no better way to tell if a market has trends than to test it for a variety of trending methods. A market that performs well using an interday breakout, but fails with a dualtrend crossover, is not a good candidate for trending profits in the future.

A comprehensive look at 12 trading methods applied to 12 futures markets gives us a chance to see generalized performance. The annualized percentage returns for a variety of U.S. markets are shown for 12 welldefined technical methods in Table 2112. These trading systems are:

- CHL, Channel breakout, using the closing price penetration of the Nday high and low
- PAR Wilder's Parabolic System, in which the stop and reverse gets closer each day
- DRM Directional Movement, a Wilder method that averages the ups and downs separately
- RNQ Range Quotient, basing a breakout on a ratio of current change to past change
- DRP Directional Parabolic, combining DRM and PAR systems

- M11 MII Price Channel, a breakout system using only the oldest and most recent prices
- LSO LSO Price Channel, using two parameters, including an interval of oldest prices, to decide a channel breakout
- REF Reference Deviation, determining a volatility breakout based on a standard deviation of past closing price changes

Louis P Lukac, B. Wade Brorsen, and Scott H. Irwin, "How to test profitability of technical trading systems Futures (October 1987).

TABLE 2112 Percentage Returns by System and Market, 1978-1984*

	CHL	PAR	DRM	RNQ	DRP	MII	LSO	REF	DA
Corn	22.4	3.8	29.8	-6.2	30.1	43.3	10.7	4.6	3
Cocoa	10.0	-101.7	-121.9	-345.0	-112.1	-73.6	-256.9	-120.5	-7
Copper	-15.4	2.9	-46.1	-78.4	-4.2	-31.4	-83.0	-66.2	-3
Cattle	-12.2	-28.4	-12.4	-72.5	-16.5	-34.8	-44.9	-56.3	-1
Pork bellies	-30.8	22.0	-6.8	-134.4	4.8	20.6	-117.4	-145.3	-
Lumber	38.7	-43.6	-0.4	-19.2	-40.6	31.3	-46.3	-18.4	3
Soybeans	7.7	-19.1	25.8	-57.9	-10.0	13.3	-21.5	-45.6	
Silver	60.5	54.4	0.2	-15.9	82.3	-19.1	12.5	-72.4	3
Sugar	103.4	46.2	61.2	-42.1	63.4	72.6	-0.8	-47.3	8
British pound	-3.5	8.1	20.7	-36.1	30.3	1.9	-7.8	3.5	2
German mark	66.5	35.4	68.2	18.8	78.0	63.3	19.2	-17.8	4
T-bills	108.7	48.5	132.7	-40.1	225.5	189.9	221.8	239.8	12
Average	29.7	2.4	12.6	-69.1	27.6	23.1	-26.2	-28.5	2

*Calculation of annual percent returns assumes that 30% of the initial investment is used for initial margin of contract value.

546

- DMC Dual Moving Average Crossover, holding a trend position when both moving averages have the same trend
- DRI Directional Indicator, a ratio of current price change to total past price change
- MAB Moving Average with % Price Band, giving a signal when prices penetrate the band
- ALX Alexander's Filter Rule, generating signals when prices reverse from a previous swing high or low by a percentage amount

To create the results in Table 2112, each market was tested for 3 years and the best parameters used to generate the returns for the next years, in a stepforward approach. It is easy to see that some markets are not profitable for any strategy and that some strategies are generally poor. Unless you were creating a strategy for a specific group of markets, you would not want to trade a system that was not profitable in less than 50% of the markets, nor would you want to trade a market that failed in most of the trend strategies. For example, cattle lost in every system, and the RNQ, REF, and MAB systems were consistently unprofitable; therefore, none of

those would be good candidates for trading. The most consistent systems were CHL, PAR, M11, and DMC, each posting 8 of 12 winning markets; yet, each has a noticeably different technique. Tbills present an interesting choice because the highest profit, 239.8%, occurred in one of the least reliable systems, REF; choosing DRP is likely to be a better choice.

Displaying the system results by year for all markets is another way to look at robustness. In Table 2113, we see that RNQ and MAB net losses over all markets in all years, while CHL, M11, and DMC stand out as very consistent. The combined presentation of results across all markets and systems is a clear way of avoiding overfitting by looking at the big picture.

PRICE SHOCKS

Price shocks are large changes in price caused by unexpected, unpredictable events. The impact of price shocks on historic tests can change the results from profits to losses, and vary the risk from small to extremely large and, unfortunately, not enough thought is given to how these moves affect test results and future performance. A discussion of price shocks could easily fill an entire book, but the concepts that are important can be explained

TABLE 2113 Percentage Returns by System and Year*

System	1978	1979	1980	1981
CHL	6.1	47.9	81.6	21.8
PAR	18.5	16.4	54.5	29.9
DRM	29.3	21.7	92.2	-31.6
RNQ	-57.2	-69.0	-9.6	-152.1
DRP	34.8	63.8	88.7	31.0
M11	12.9	41.6	87.8	54.6
LSO	-47.1	-4.0	39.1	-55.1
REF	-13.5	23.3	53.2	-85.2
DMC	17.6	26.8	85.4	5.9
DRI	-60.7	-15.4	46.3	-108.1
MAB	-38.0	-44.3	-7.7	-114.4
ALX	28.3	38.6	82.6	-34.4
Average	-5.8	12.3	57.8	-36.5

*Equal percentage amounts of margin are assumed to be invested in each market.

briefly. When it comes to actual trading, the difference between your expected results and actual performance (not attributed to slippage) is entirely dependent on the number of price shocks.

By its very nature, a price shock must be unexpected. However, not all events cause shocks. An assassination, abduction, or political coup is

likely to be a surprise. while a crop freeze can be anticipated as weather turns unusually cold. Some economic reports, such as a .5% increase in the Producer Price Index or jump in the balance of trade, will come as a shock, but a low carryover supply of soybeans or a tightening of the money supply after steady economic growth can be anticipated. When a change is anticipated, the market adjusts to the correct level before the news is announced, when it is wrong, prices react in proportion to how poorly it was anticipated.

The problem comes when you test historic data. We know at the time of a price shock that we could not have anticipated the event; at best we have an even chance of being on the right side of the price move. But the computer doesn't know that, and the best results from a large series of tests often include profiting from the largest price shocks a situation that would be impossible in actual trading. For example, if there were 10 price shocks in 10 years of data, and your historic tests profited from 8 of those shocks, then you have overestimated your profits. Even worse, you have underestimated your risk by believing that a price shock produced a profit and not a loss.

How can you correct this problem? First, you can look at the pattern of profits and losses from a series of tests. If a 20day and 30day moving average each showed a maximum drawdown of \$10,000, but the 25day average only lost \$5,000 you should assume that the 25day results could have had the same drawdown as the nearby tests. By some quirk of timing one test was able to avoid a loss while the normal test was not. It would be risky to assume such good fortune in trading.

If you have a record of major price shocks, such as a chart published by the exchanges showing key events, you can verify how many of these events produced profits or losses in your final system. If you have profited from more than 50% then you should reassess your risk based on losses rather than profits from some of those trades. Also check to see that there is an even distribution with regard to the size of the price shocks. You may have selected parameters that took small losses on minor shocks and large profits on major ones.

If you do not treat price shocks as a serious risk, you may overleverage and undercapitalize your trading account. This is the most common cause of catastrophic loss. Any price shock that causes a windfall profit could have, just as easily, produced a devastating loss. The example that follows, 'Anatomy of an Optimization,' shows how expectations and reality are often quite different. Further examples can be found in Chapter 22 ("Practical Considerations").

ANATOMY OF AN OPTIMIZATION

Using the Iraqi invasion of Kuwait on August 6, 1990, and the U.S. retaliation on January 17, 1991, we optimized a simple moving average strategy for crude oil. The test period began January 2, 1990, but 100 days were needed to initialize the calculations; therefore, the first trade was entered May 24, 1990 (see Tables 2114, 2115, and 2116). In both Test 1 and Test 2, the optimization selects moving averages that produced the highest profits. By accepting these choices, the trader would have held a

short position before the invasion of Kuwait, and a long position before the U.S. retaliation. In reality, there would have been a large loss rather than an exceptional profit. When a price shock is an important part of the data being tested, the best system is the one that took the most profits out of that move, even at the cost of other trades.

"Perry Kaufman, "Price Shocks: Reevaluating Risk/Return Expectations," Futures Industry (june/july 1995).

TABLE 21.14 Test 1: Optimizing Crude Oil, Jan through August 3, 1990*

<i>Period</i>	<i>NetPrft</i>	<i>ROA</i>	<i>Max</i>
5	2615	82.9	-31
10	6765	959.6	-7
15	6975	989.4	-7
20	4975	705.7	-7
25	6975	989.4	-7
30	4635	657.5	-7
35	3635	313.4	-11
40	3215	276.0	-11
45	3255	187.6	-17
50	255	10.6	-24
55	-715	-21.1	-33
60	2625	156.7	-16
65	2785	239.1	-11
70	-385	-12.6	-30
75	1955	112.7	-17
80	-5040	-100.0	-50
85	-5040	-100.0	-50
90	-5040	-100.0	-50
95	-5040	-100.0	-50
100	-5040	-100.0	-50

Trade Detail from Testing

<i>Trade Date</i>	<i>B/S</i>	<i>Price</i>	<i>Pr</i>
May 24, 90	Buy	23.58	
May 25, 90	Sell	23.55	
Jul 11, 90	Buy	22.49	10
Aug 3, 90	Sell	28.51	50

* Testing stopped the day before the invasion of Kuwait.

Test selects 15-day moving average.

Holding short on August 6, 1990, when Iraq invades Kuwait.

in Test 2, although there were large profits from the first price shock, the trader would have still held the wrong position when the second shock

occurred. This method of testing hides the real trading risk, focuses on profits that cannot be predicted, and does it all at the expense of consistent trading profits.

DATA MINING AND OVEROPTIMIZATION

Advances in computing power and testing software have made it very easy to test trading strategies using historic data. It is only logical that you should prove that your ideas would have worked in the past. It is also sensible that you understand the amount of risk that you must take to gain the results. For more sophisticated analysts, a relationship may be found between fundamental factors, volatility, and risk. For fundamental analysts, the expected price of oil is greatly dependent upon the crude inventories., for many stock issues, the per capita disposable income is the key element in predicting revenues; and, the anticipated volatility of soybeans is often found in the size of the carryover stocks. Although past prices

TABLE 21-15 Test 2: January 2, 1990, through January 16, 1991*

<i>Period</i>	<i>NetPrft</i>	<i>ROA</i>	<i>M</i>
5	-6010	-32.3	-
10	3660	22.3	-
15	12350	103.8	-
20	19810	345.4	-
25	8410	105.5	-
30	8420	86.8	-
35	18075	1348.9	-
40	11150	308.9	-
45	9335	304.1	-
50	8265	246.7	-
55	2485	55.0	-
60	9075	184.8	-
65	7155	100.4	-
70	3815	53.3	-
75	7195	83.9	-
80	-8540	-82.6	-
85	-3860	-54.2	-
90	1960	35.7	-
95	440	9.5	-
100	-3950	-85.6	-

Trade Detail from Testing

<i>Trade Date</i>	<i>B/S</i>	<i>Price</i>
May 24, 90	Buy	23.58
May 25, 90	Sell	23.55
Jul 12, 90	Buy	23.49
Oct 19, 90	Sell	40.10
Oct 30, 90	Buy	41.23
Oct 31, 90	Sell	41.92
Nov 19, 90	Buy	38.39
Nov 21, 90	Sell	37.30
Nov 26, 90	Buy	40.62
Nov 27, 90	Sell	40.53
Jan 7, 91	Buy	35.70

* Testing stopped the day before the U.S. retaliation.

Test selects 20-day moving average.

Total profits are \$3,225 without price shock.

Holding long when U.S. attacks Iraq.

and their relationships to economic data cannot be ignored, they can also be misleading, or even erroneous, if used carelessly.

The most important factor in the success of a trading program is a sound premise; that is, you must find a real-life relationship between the price and the way participants react. For example, when the monetary authority of any country raises interest rates, the price of equities tends to fall. Or, when there is bad news for the economy, investors shift from equities to guaranteed interest rates, driving the stock market lower and the

bond market higher. Once these relationships have been identified, it is perfectly sensible to look

TABLE 21-16 Test 3: January 2, 1990, through March 28, 1991*

<i>Period</i>	<i>NetPrft</i>	<i>ROA</i>	<i>MaxDD</i>	<i>Trds</i>
5	-17255	64.0	-26975	45
10	-13195	-45.0	-29295	31
15	4480	26.6	-16840	20
20	8430	67.6	-12475	16
25	-4690	-22.6	-20725	20
30	-1395	-6.4	-21775	17
35	28335	2114.6	-1340	9
40	-360	-2.5	-14420	12
45	18245	594.3	-3070	7
50	8265	246.7	-3350	3
55	2485	55.0	-4515	7
60	9075	184.8	-4910	5
65	7155	100.4	-7130	7
70	3815	53.3	-7160	7
75	7195	83.9	-8580	9
80	-18225	-81.3	-22410	13
85	-12945	-70.4	-18390	9
90	-4065	-40.3	-10080	5
95	-8645	-56.0	-15430	5
100	-3485	-75.5	-4615	5

Trade Detail from Testing

<i>Trade Date</i>	<i>B/S</i>	<i>Price</i>	<i>Profit/ Loss</i>	<i>Cum P/L</i>
May 24, 90	Buy	23.58		
May 25, 90	Sell	23.55	-55	-55
Jul 16, 90	Buy	24.16	-635	-690
Nov 9, 90	Sell	40.58	16395	15705
Jan 14, 91	Buy	38.76	1795	17500
Jan 16, 91	Sell	39.36	575	18075
Mar 8, 91	Buy	29.89	9445	27520
Mar 13, 90	Sell	30.89	975	28495
Mar 19, 90	Buy	30.75	115	28610
Mar 28, 90	Sell	30.50	-275	28335

* Test period includes both price shocks.

Test selects 35-day moving average.

Total profits are \$2,495 without price shocks.

Optimization profits include both Iraqi invasion of Kuwait and the U.S. attack on Iraq.

through historic data to uncover the size of the interest rate increase needed to move the market, or the type of pattern shift from equities to bonds that

could yield a profitable trade.

When we begin looking for a relationship between inventories and price we are reasonably confident that there is a solution, and we can even identify those elements that form the solution. The process is one of careful reasoning and a small amount of testing, or validation. However, when we test a set of trading rules on historic data, we do not actually know that this method will yield a profitable result.

551

For example, take the simplest case of a trading system that only uses a moving average trendline. A long position is signaled when the trendline turns up, and a short is generated when the trendline turns down. Is this a sound premise? How do you decide that a profitable test result will become profitable realtime trading? The process for understanding this is far from trivial.

Realistic Assumptions

After you have decided on your underlying strategy, and before you begin testing, there are a number of decisions to make. These must be made without the benefit of hindsight, but according to your own reasoning. They include:

1. The amount of data to be tested
2. The commission cost per unit traded
3. The slippage cost per unit traded
4. The range of each parameter (variable) to be tested

This last item, the range of each parameter, means that you must decide, in advance, which moving average periods will be tested. If this is to be a longterm program trying to capitalize on the way governments manipulate interest rates or money supply, then you might choose periods beginning with 50 days, up to 300 days. If you believe that the longterm is too risky and most trends are mediumterm, then a range of 20 to 100 days might be best. The very fast day trader, looking to downplay the trend, would want to test periods from 5 minutes to 1 day. It is important that the periods

chosen for testing correspond to how you perceive the price movement. It is not enough to say, "I'll trade anything that makes money." Commissions and slippage are also vital to the results, but more significant for the shortterm trader. As a system focuses on shorter holding periods, the profits per trade must become smaller; therefore, the commissions and slippage can become the difference between profits and losses. Because a computer assumes that orders can be executed perfectly (at the signal price), even during a 5minute bar where prices jumped 1% with virtually no trading, it takes some effort to decide on a realistic level of slippage. Normally, it is safe to overestimate the cost of entering and exiting the market, but for a shortterm trader, even a small overestimation could eliminate a potentially profitable trading technique. Accuracy, as near as possible, is the best approach.

By making clear assumptions in advance, you gain the advantage of validating your ideas. You thought that mediumspeed trends were likely to be best, but test results show that only a narrow range of speeds are profitable. You must now rethink why this occurred. Using negative

feedback to learn more about the market is a very healthy process.

Statistics Can Say Whatever You Want

Most computer software that allow historic testing give results that emphasize the highest returns. Given a choice, most users of these programs want to see the highest returns. A single test result, however, should not be considered representative of the worth of the system being tested.

If we test the same 10year period of historic data for 1,000 different parameter values, spread over a wide range, we are likely to be pleased if 50 of those test results were very attractive. Statistically, we might say that 50 are significant at the 5% level. But that is not necessarily true. In a welldistribution, or random, set of results, we can expect 5% to be statistically significant. If the trading system is truly sound, then we would like to see much more than 5% successes.

552

Robustness

Practically speaking, robustness translates into getting the most realistic results by using the fewest parameters and the most data. Robustness is a term used to describe a system or method that works under many market conditions, one in which we have confidence; it is associated with a successful result using an arbitrary set of parameters for a system. This is most likely to occur when only a few parameters are tested over a long time period. For example, if you claim that a longterm moving average system is robust, then someone unfamiliar with the system should be able to choose the moving average period of 130 days and have reasonable expectations of a profit. This would be most likely if a large percentage of all time intervals in the longterm test were profitable.

When first testing a system, it is best to avoid looking at the largest profits, but instead, study only the average values of all tests, that is, the average net profit, average number of trades, average drawdown, and average profit per trade. in addition, looking at the standard deviation of profits for all tests will give a very good idea of the consistency of returns over the range of tests.

The following is a checklist along with a brief explanation that might help when trying to follow a procedure to get robust test results:~

1. Deciding what to test. Testing should be used for validation of an existing idea, not for discovery.
 - a. Is the strategy logical? it should be based on careful observation of the market, or an awareness of relationships between different factors that drive the market, or between different markets themselves.
 - b. Can you program all the rules? Programming a system guarantees that all of the contingencies have been thought out. While not everything can be programmed, in particular situations requiring crisis intervention, these cases are few.
 - c. Does the strategy make sense only under certain conditions? A trading program can be limited to a specific situation, and testing should reflect those limitations.
 - d. Take a guess as to the expected results. It is important to know if your

estimate of the system success is correct. Getting a good result for the wrong reason requires just as much rethinking of the problem as getting bad test results.

2. Deciding how to test. There are a number of steps necessary before you can actually begin the test process.

- a. Choose the testing tools and methods. Select a test platform, such as TradeStation or MetaStock, which can speed up the test process and keep it objective.
- b. Do you have enough of the right data? More data is best, but if there is less data available, it should have a wide range of market conditions, such as bull and bear markets, price shocks, and sustained sideways periods.
- c. Will you test a full range of parameters? The nature of your trading strategy, for example, longterm or shortterm breakouts, may determine the parameters that are tested. Test only the range that makes sense for the strategy. This serves to validate the concept and avoids discovering better techniques.
- d. In what order will the parameters be tested? Test the most important parameters first, such as the time period. More subtle parameters are often used for refining the results, and may not be helpful at all. You can usually tell the most important parameters because they cause the widest range of results.
- e. Are the parameters distributed properly? Testing equal intervals of trend periods can be misleading. The difference between a 3 and 4day moving average is

A more detailed explanation of testing robustness can be found in Perry Kauffman, *Smarter Trading* (McGrawHill, 1995).

553

33%, while the difference between a 99 and looday average is only 1%.

When you test periods from 1 to 100 days, your average results are heavily weighted toward the slow end, which has similar performance.

- f Have you defined the evaluation criteria? You must know whether highest profits, highest reliability, or some other combination will be used to evaluate the results. Riskadjusted performance can be a very good objective criteria.
 - g. How will the output be presented? Results can look very different when presented in a plain table. Two and 3dimensional contour maps can give a better view of the way performance varies when two parameters change.
3. Evaluating the results. Careful review of the test results can avoid many hours of unnecessary work. Looking at the detail of the performance, and even the individual trades of a few selected results, will give you insight into the good and bad aspects of your program.
- a. Are the calculations correct? just because calculations were done by a computer doesn't mean they are always correct. A person had to tell the computer what to do, and people make mistakes.
 - b. Were there enough trades to be significant? We have uncertainty when

the results show excellent performance on only a few trades. When there is not enough data, there is no way to be sure that the results are robust. The system can only be used if you have complete confidence in the underlying premise, and the sparse results simply confirm that logic.

- c. Does the trading "stem produce profits for most combinations of parameters? The best assurance of robustness is when most parameter combinations produce profitable results, even though they vary.
 - d. Did logic changes improve overall test performance? When you improve the results by adding a special rule, the average of all tests must increase, while the standard deviation remains unchanged. Otherwise, you may have simply overfit part of the data.
 - e. How did it perform on outofsample data?
4. Choosing the specific parameters to trade. Choosing the parameters to be used in trading is another unique and difficult problem. Does the system that showed the peak historic profits give you the best chance of getting the peak future profits, or was it just chance?
- a. Did the last test include the most recent data? If you have left out data to perform an outofsample test, then respecify the parameters using the most recent data. You must be as current as possible before trading.
 - b. Did you choose from a broad area of success? Parameters should be chosen from an area that showed consistent profits, but should not be determined by the one that had peak profits. It is best to look for average returns using average parameters.
 - c. Are profits distributed relatively evenly over the tested history? Although there may be attractive total profits, sustained periods of loss would not survive real trading. Returns should be reasonably steady.
 - d. Did the historic results show any large losses due to price shocks? it is not possible to avoid large price shocks in real trading because you never know when they will come. Historic results that do not show large equity jumps due to price shocks are not realistic, and result from overfitting. Even worse, those results that only profit from price shocks are benefiting from hindsight. In real trading, 50% of the price shocks should produce a windfall profit, while the other 50% will give equal losses.

554

e. Have you riskadjusted the returns to your acceptable risk level? The best comparison between tests is to annualize all returns, to avoid comparing unequal time periods, and to riskadjust the results. Risk adjusting changes each result so that it has the same unit risk as every other test.

5. Trading and performance monitoring. The development of a system is never complete. There is a continuous evolution in the markets, seen in the changing participation, volatility, interrelationships between markets, and the introduction of new markets. The only way to see

how these changes affect your trading program is to carefully monitor the system signals and the corresponding fills. From this evaluation you can reallocate funds, change the way you execute signals, and even find better rules. Without monitoring your trading, and comparing the results with your expectations, there is no basis for saying that your system is performing correctly.

- a. Are you following the same rules that were tested? It is often convenient to program a simple version of your ideas, but trade a more complex set of rules. If so, you have no basis for comparison.
- b. Are you trading the same data that was tested? The use of an artificial, continuous contract for testing futures will not give you the same results as testing each contract, beginning at the time it becomes the most active of the delivery months. Normally, artificial continuous prices are smoother than real prices and appear to perform better.
- c. Are you monitoring the difference between the system and actual entries and exits? Slippage, the difference between the system price and the actual market execution, can change performance from theoretically profitable to a real trading loss. Monitoring this difference will tell you the realistic execution costs, the maximum volume that can be traded, and the best time of day to trade.

SUMMARY

This chapter has covered a lot of material vital to the successful development of a trading program. Because there are so many issues, the following is a brief summary of the most important concepts:

1. Select a trading strategy with a sound underlying premise; complexity is not necessary.
2. Reserve out-of-sample data for validation.
3. Select the range of parameters to test that is logical, favoring the faster or slower trading approach.
4. Perform the longest reasonable test to include as many market situations as possible.
5. Test as many different markets as possible looking for a common pattern in the results.
6. Standardize the results to facilitate comparisons between systems and test periods.
7. Evaluate the robustness of the method rather than the peak performance; avoid fine tuning.
8. Assess the effects of past price shocks on profits and risk.
9. Bet-test (paper trade) the system until you are certain that it is working properly.

You might also keep in mind that testing uses historic data and works best in real trading when the current market has the same patterns seen during the test data. The premise of optimization is that the markets will continue to behave in the future as they have in the past. This may be why the index markets traditionally have been among the poorest system performers because they are trading at price levels and volatility that have never been seen before.

other areas of interest to readers actively testing and evaluating test performance will be genetic algorithms, a new search technique, discussed in Chapter 20, and most of Chapter 23, especially the section "Measuring Risk."